

**DISCOVERING ALTERNATIVE SPLICING OF DEEPLY
CONSERVED EXONS AND CHARACTERIZING THE
INTRONOME IN THE UNICELLULAR YEAST *S. POMBE* USING
LARIAT SEQUENCING**

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy
By
Ali Raza Awan
August 2013

© 2013 Ali Raza Awan

**Discovering Alternative Splicing of Deeply Conserved Exons and
Characterizing the Intronome in the Unicellular Yeast *S. pombe* using
Lariat Sequencing
Ali Raza Awan, Ph.D.
Cornell University 2013**

Alternative splicing is a potent regulator of gene expression that vastly increases proteomic diversity in multi-cellular eukaryotes. Although it is widespread in vertebrates, little is known about the evolutionary origins of this process owing in part to the absence of phylogenetically conserved events that cross major eukaryotic clades. The unicellular fission yeast, *Schizosaccharomyces pombe*, is an organism evolutionarily distant from mammals that nonetheless shares many of the hallmarks of the major mammalian alternative splicing form, exon skipping. Further, *S. pombe* is a highly genetically tractable organism that has been considered as an attractive potential model system in which to study exon skipping. However, evidence of exon skipping from RNA-seq studies in *S. pombe* has remained elusive.

To better search for such evidence I have developed a novel lariat sequencing approach that offers high sensitivity for detecting splicing events, and applied it to study splicing in *S. pombe*. Using this approach, I discovered multiple examples of exon skipping, several of which involve exons that are conserved with dozens of animals, plants, other fungi and even protists. Strikingly, some of the specific alternative splicing patterns found in *S. pombe* are identical in mammals. In addition, I discovered hundreds of novel splicing events, many of which occur in genes that were thought to be intronless. Finally, I present only the second large-scale map of the intronic branch point sequences in any organism, allowing me to test a prediction about differences in splice site recognition in introns from non-coding regions versus those that separate protein-coding exons.

BIOGRAPHICAL SKETCH

Ali Raza Awan was born on June 29th, 1981 to Munia Rab and Kazim Awan. Ali was born in London, but moved with his family to Pakistan at the age of six months. After three years in his father's family home in Quetta, Ali moved again, this time to Jeddah, Saudi Arabia, where he spent most of his childhood. It was in Jeddah that Ali developed his love for nature and the outdoors. Situated next to the Red Sea and near the Empty Quarter, Jeddah was ideally situated for the young Ali to explore the riches of the beach as well as the vastness of the desert. Ali spent many a happy weekend with his parents, his brother and family friends picnicking by the seaside, or on trips out into nearby mountains. It was during this time, snorkelling in the Red Sea, collecting seashells and hunting for fossils, that Ali became passionate about the natural world.

After moving to Dubai as a teenager, Ali was exposed to new aspects of the science of biology. At the beginning of high school, Ali's father bought him a book called "Introducing Genetics" by Steve Jones and Borin Van Loom. Little did he know that this book would propel Ali on a path to his future career. Genetics combined Ali's love of nature with his fondness for mathematics (his mother taught the subject). The Biology A-levels in high

school formalized Ali's foray into genetics, which was further bolstered by an inspirational biology teacher, Mr Lochhead.

After high school, Ali was accepted into Cornell University as an undergraduate, where he began his college career majoring in biology and computer science. On a whim, sometime during the winter of his sophomore year, Ali decided to apply to transfer to Stanford University. Six months later, Ali made the switch and spent three wonderful years soaking up the California sunshine, pursuing a Bachelors degree in Biology and a Masters degree in Bioinformatics, growing his hair, playing the guitar and becoming a hippie of sorts. After graduating and two brief stints at biotech start-ups in St. Louis and Dubai, Ali decided that it was time to get a PhD. Funnily enough, he ended up back where he started: at Cornell University, with many cold winters ahead of him. At Cornell, Ali decided that the lab of newly appointed Assistant Professor Jeff Pleiss was the perfect academic home. The research promised an exciting mix of genomics and bioinformatics where Ali felt that his background would nicely complement the high-throughput methods-based outlook of the lab. When he is not working on science, Ali is interested in music (listening and playing), literature, football (soccer), and spending time with family and friends.

ACKNOWLEDGEMENTS

I thank my parents Munia and Kazim and my brother Taimur, for all of their love and support, without which none of this would have been possible. This was especially important during some of the times when it felt like winter would never end in Ithaca.

I thank Jeff for sharing his immense enthusiasm for doing science, and showing me how to be more positive about my own work. I also thank Jeff for teaching me the importance of effective communication in science, right from my very first department seminar down to the time I presented my work in the RNA Society conference in Kyoto.

I owe Eric Alani a huge debt of gratitude for always having an open door and giving me sound advice about how to navigate the challenges of the PhD during some of the times when Jeff unfortunately had to be away. I extend thanks to Andrew Grimson for the similar role he played, and for the encouragement he gave me. I thank John Lis for the support and advice he has given me along the way, especially with regard to my choice of postdoctoral position.

I thank my friends, especially Satyaki Prasad, Syud Ahmed, Gabriel Hoffman, Yin He and Stephane Bentolila for all the good cheer through these six years and just for making Ithaca a lot more interesting in general.

Finally, I am grateful to my aunt Bugs Khala and my mum (again) for putting me up and putting up with me to make it possible to write my thesis during an intense two week period in the Upper West side of Manhattan.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

| | |
|--|----|
| 1.1 Alternative Splicing is a Key Modulator of Eukaryotic Gene Expression | 1 |
| 1.2 The Regulation of Alternative Splicing is Multi-Layered | 11 |
| 1.3 Genome-wide Approaches to Study Alternative Splicing in a Genetically Tractable System | 19 |

CHAPTER 2: DEVELOPING LARIAT SEQUENCING TO DETECT EXON SKIPPING IN *S. POMBE*

| | |
|---------------------------------|----|
| 2.1 Introduction | 40 |
| 2.2 Results | 43 |
| 2.3 Discussion | 78 |
| 2.4 Materials and Methods | 83 |

CHAPTER 3: OPTIMIZING LARIAT SEQUENCING TO BETTER PROFILE THE WHOLE *S. POMBE* INTRONOME

| | |
|---------------------------------|-----|
| 3.1 Introduction | 121 |
| 3.2 Results | 123 |
| 3.3 Discussion | 145 |
| 3.4 Materials and Methods | 146 |

CHAPTER 4: MAPPING *S. POMBE* BRANCHPOINTS ON A GENOMIC SCALE

| | |
|------------------------|-----|
| 4.1 Introduction | 176 |
| 4.2 Results | 180 |
| 4.3 Methods | 195 |

CHAPTER 5: FUTURE DIRECTIONS

| | |
|-------------------------------------|-----|
| 5.1 Srrm1 and Exon Definition | 207 |
| 5.2 Circular Exonic RNAs | 217 |

CHAPTER 1: INTRODUCTION

1.1 Splicing is a Key Modulator of Eukaryotic Gene Expression

The Nobel Prize-winning discovery of “split genes” in the late 1970’s (Berget, Moore, & Sharp, 1977; Chow, Gelinas, Broker, & Roberts, 1977) dramatically changed our understanding of the architecture of eukaryotic gene structure. While it was already known that non-coding untranslated regions flanked the protein-coding segments of genes, this discovery led to the idea that the protein-coding region itself consisted of coding segments (exons) interrupted by non-coding regions called introns (Cech, 1983). Intronic sequences are typically removed from the RNA transcript in the nucleus via a process called splicing (Padgett, Grabowski, Konarska, Seiler, & Sharp, 1986).

The chemical mechanism of splicing consists of two sequential transesterification reactions ((Konarska & Query, 2005; Staley & Guthrie, 1998), Figure 1.1). In the first reaction, the 2’ hydroxyl of an adenosine residue upstream of the 3’ end of the intron undertakes a nucleophilic attack on the phosphate group at the three prime end of the upstream exon. This reaction leaves a free hydroxyl group at the 3’ end of the upstream exon, as well as a circularized intron followed by the downstream exon, collectively termed the “lariat intermediate” (Madhani & Guthrie, 1994). The lariat intermediate

contains an unusual 2'-5' phosphodiester linkage between the internal (branch-point) adenosine and the guanosine at the 5' end of the intron (Figure 1.1). In the second transesterification reaction, the free 3' hydroxyl group at the 3' end of the upstream exon undertakes a nucleophilic attack on the phosphate group at the 3' end of the intron, resulting in ligation of the upstream and downstream exons via canonical 3'-5' phosphodiester linkage, as well as the excision of the circular (lariat) intron species ((Staley & Guthrie, 1998), Figure 1.1). The chemistry of splicing is performed by the spliceosome, a highly dynamic ribonucleoprotein machine that forms anew on every intron (Madhani & Guthrie, 1994; Wahl, Will, & Lührmann, 2009). The spliceosome is comprised of five highly conserved small nuclear RNAs (snRNAs) – U1, U2, U4, U5 and U6 – as well as over one hundred proteins (Bessonov, Anokhina, Will, Urlaub, & Lührmann, 2008; Jurica, Licklider, Gygi, Grigorieff, & Moore, 2002), most of which are associated with the snRNAs as ribonucleoprotein complexes (snRNPs).

During splicing, the snRNAs associate with intronic sequence elements by base pairing, in the following order. First, the U1 snRNP associates with a sequence element at the 5' end of the intron, called the 5' splice site (5'ss). Subsequently, the U2 snRNP associates with the sequence surrounding the branch point adenosine, called the branch point sequence. Next, the remaining three snRNPs are recruited to the intron in the form of a “tri-snRNP”:

U4/U6.U5, where U4 and U6 interact with each other via base pairing. A series of RNA rearrangements mediated by RNA helicases then displace U4 and U1, rendering the spliceosome catalytically active (Nilsen, 1998). Following the completion of splicing chemistry, spliceosomal components dissociate from the RNA and are recycled to be able to participate in further rounds of splicing (Wahl et al., 2009).

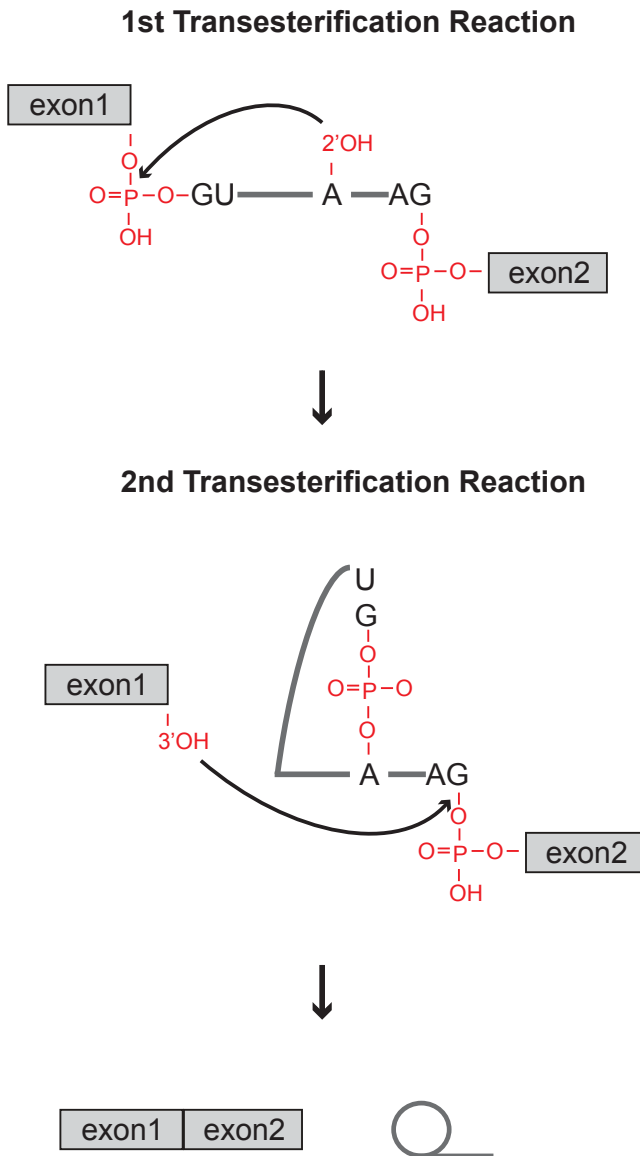


Figure 1.1

The chemistry of splicing. The first and second transesterification reactions that take an unspliced RNA to the mature exon-exon ligated RNA as well as the excised lariat intron are depicted.

Since introns are typically excised prior to translation they do not directly affect the sequence of the translated protein (Padgett et al., 1986). However, the presence and subsequent removal of introns in a gene has been shown to positively affect gene expression at the levels of mRNA export and translational efficiency (Nott, Le Hir, & Moore, 2004; NOTT, MEISLIN, & MOORE, 2003). Further, the presence of introns can also negatively affect gene expression. The mere presence of intronic sequences does not guarantee splicing; splicing is a sequential process with multiple points at which it can be disrupted (Guthrie, 1994; Wahl, Will, & Lührmann, 2009). The failure to remove an intron often introduces premature termination codons in the mRNA that is exported for translation. This can happen either by the presence of an in-frame termination codon in the sequence of the intron itself (Pollard, Flanagan, Newton, & Johnson, 1998; Senapathy, 1986) or because the intron length is not a multiple of three, such that its inclusion changes the reading frame, which can lead to new in-frame termination codons (Carbone, Applegarth, & Robinson, 2002; Hayashi, Gekka, Omoto, Takeuchi, & Kitahara, 2005). Hence, the failure to remove these introns by splicing can lead to either the production of non-functional truncated proteins (Michael et al., 2005), or to the degradation of the pre-mRNA precursor itself via a process called

nonsense-mediated decay (Hentze & Kulozik, 1999; Maquat, 1995). Such a paradigm of splicing-mediated quantitative regulation of gene expression has been shown to play an important role in cellular responses to stress (Pleiss, Whitworth, Bergkessel, & Guthrie, 2007) and during development (Anastasaki, Longman, Capper, Patton, & Cáceres, 2011; Wittkopp et al., 2009). These are examples of the production of different mRNA transcripts via modulation of splicing, a process that is referred to as alternative splicing.

In addition to quantitative regulation of gene expression, alternative splicing plays an arguably greater role in the qualitative regulation of gene expression. In the cases mentioned above, failure to remove an intron via splicing resulted in premature termination codons, leading to a non-functional protein or pre-mRNA degradation. However, often the intronic sequence contains no in-frame termination codons and its length is divisible by three, such that its inclusion in the processed transcript leads to a translatable mRNA that encodes a functional protein product (Riccombeni, Vidanes, Proux-Wéra, Wolfe, & Butler, 2012; Strijbis, den Burg, F Visser, den Berg, & Distel, 2012). Such “productive” alternative splicing allows a single gene to code for multiple functional protein isoforms. The differential inclusion of introns in mRNA by splicing is only one type of alternative splicing, referred to as “intron retention” (Blencowe, 2006). Intron retention can be described as a decision by the splicing machinery about which intronic splice sites to use: by using or not

using the 5'ss-3'ss pair that define the boundaries of an intron, the splicing machinery either causes excision of that intron or leaves it in the processed mRNA (Figure 1.2a).

Whenever a gene contains more than a single intron, the pairing of a 5'ss and a 3'ss from *different* introns by the splicing machinery allows the differential inclusion of exons rather than introns in the processed transcript (Figure 1.2b). This type of alternative splicing is referred to as “exon skipping” (Blencowe, 2006).

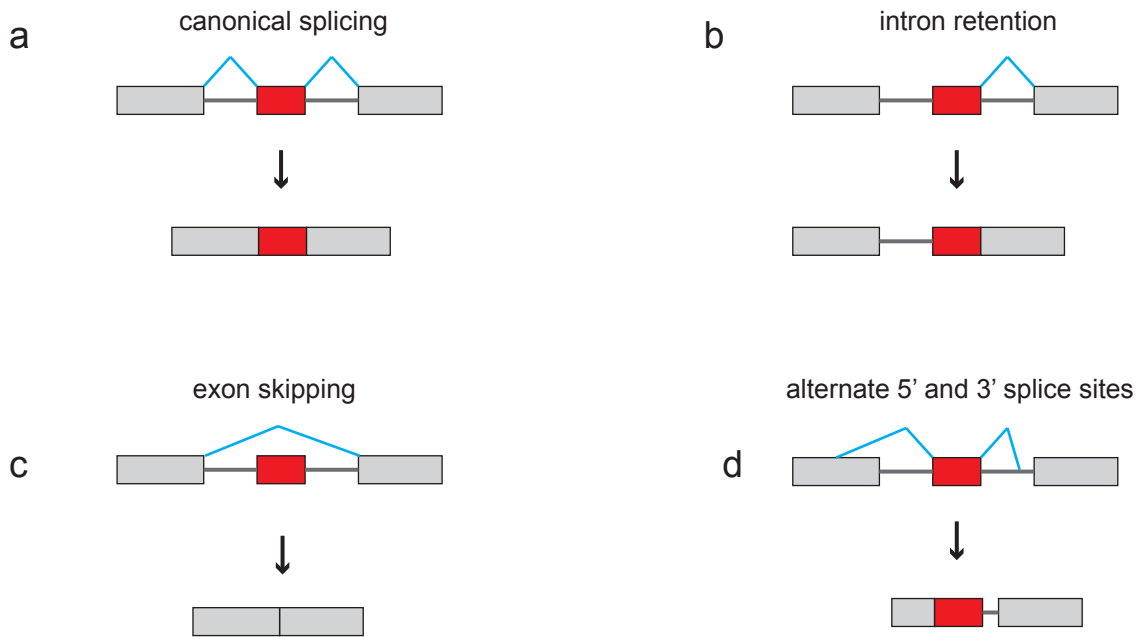


Figure 1.2

Major Types of Alternative Splicing. (A) normal splice pattern where introns 1 and 2 are removed by splicing. Grey lines represent introns, grey and red boxes represent exons. Blue lines indicate splicing pattern. (B) Intron retention, where intron 2 is removed by splicing but intron 1 is not (C). Exon skipping, where splicing excises exon 2 as well as introns 1 and 2. (D) Alternative 5' and 3' splice sites, where splicing results in different boundaries for introns 1 and 2, which are both removed.

As the number of introns in a gene increases, the number of mathematically possible combinations of different non-terminal exons in the processed mRNA, and hence different proteins, increases exponentially. This observation has been thought to contribute greatly to the expansion of protein coding capacity in multi-cellular eukaryotes (Maniatis & Tasic, 2002; Nilsen & Graveley, 2010). Indeed, when the human genome was first sequenced, the number of protein coding genes inferred from the sequence was far lower than was expected, given the number of proteins that were known (Collins, Lander, Rogers, Waterston, & Conso, 2004; Stein, 2004). Further, this number, which was under 25,000, was only marginally greater than the number of protein coding genes -- ~20,000 -- found in *C. elegans*, a far less complex organism (Claverie, 2001). It has been postulated that alternative splicing, and exon skipping in particular can largely account for the discrepancy between the number of human protein coding genes and proteins, and can to some degree bridge the gap between organismal complexity and number of protein coding genes (Graveley, 2001). The former claim has been supported in recent years by work showing that >90% of all human genes undergo some form of alternative splicing (Wang et al., 2008). Support for the latter claim was recently provided by the demonstration that protein-protein interactions between key regulatory proteins at the hubs of gene regulatory networks are likely modulated by

alternative splicing in different vertebrates, allowing significant expansion of gene regulatory potential (Barbosa-Morais et al., 2012).

At the level of individual proteins, exon skipping provides a route for modularization of protein function via the linking of key protein domains to individual exons (Franzin, Yu, Thai, Choi, & Marassi, 2005; De Roos, 2007; Torahiko et al., 1994). Alternative splicing of these exons can result in changes in protein interaction partners as mentioned above, in changes in cellular localization (Clements, Mercer, Paterno, & Gillespie, 2012; Terenzi & Ladd, 2010) and in enzymatic activity (Goulet, Gauvin, Boisvenue, & Côté, 2007; Papadopoulos et al., 2011). The phenotypic effects of exon skipping often have a switch-like logic, where the protein isoforms with or without the skipped exon can have antagonistic effects that operate at different levels of organismal biology. Examples include apoptosis, at the level of the cell (Boise et al., 1993; Schulze-Osthoff, Ferrari, Los, Wesselborg, & Peter, 1998), axon guidance, at the level of the organ (Hattori et al., 2007; Hattori, Millard, Wojtowicz, & Zipursky, 2008), and sex determination, at the level of the whole organism (Baker, 1989; Schutt & Nothiger, 2000). Given the wide ranging phenotypic effects of alternative splicing and the implication of mis-regulated alternative splicing in many human diseases (David & Manley, 2010; Singh & Cooper, 2012; Spitali & Aartsma-Rus, 2012), it is perhaps unsurprising that alternative splicing is itself subject to multi-layered regulation.

1.2 The Regulation of Alternative Splicing is Multi-Layered

Alternative splicing can be conceptualized as a series of decisions about how to pair possible five prime and three prime splice sites in a nascent transcript, leading to removal of the sequences between the paired splice sites. Several genetic and epigenetic factors are thought to play a role in making these decisions (Black, 2003; Luco, Allo, Schor, Kornblihtt, & Misteli, 2011). Both classes can be divided into *cis* and *trans* acting factors. Among the genetic factors, the most obvious *cis* elements that allow differential choice of splice sites are the splice site sequences themselves. Prior to the excision of intronic sequence during splicing, the splicing machinery (the spliceosome) is assembled on the intron in a sequential manner. To commit a pair of potential splice site sequences to splicing, the RNA components of two of the major spliceosomal components, the U1 small ribonuclear protein (snRNP) and U2 snRNP, recognize the 5'ss and key sequence elements upstream of the 3'ss called the branch point and polypyrimidine tract by base pairing (Guthrie, 1994; Madhani & Guthrie, 1994) (Figure 1.3 a). The degree of sequence homology between these components of the spliceosome and the intronic sequence elements they bind to is a major determinant of the strength of the interaction, and has allowed researchers to classify splice sites as “strong” or “weak” (Roca, Sachidanandam, & Krainer, 2005) (Figure 1.3 a). Whereas strong splice sites

tend to be constitutively used for splicing, the presence of additional trans-acting protein factors that are not part of the core splicing machinery is often required for weaker splice sites to be used (Bi, Xia, Li, Zhang, & Li, 2005; Puig, Gottschalk, Fabrizio, & Séraphin, 1999). By changing the levels of these trans-acting protein factors available to the splicing machinery, the use of alternative splicing patterns that involve weaker splice sites can also be regulated.

The best studied family of RNA-binding proteins implicated in the regulation of splice site choice is the SR proteins, so named for the long stretches of arginine and serine amino acids they contain (Shepard & Hertel, 2009; Zahler, Lane, Stolk, & Roth, 1992). During splicing, the major role of SR proteins is to bind to certain exonic sequences via an RNA-recognition motif (RRM) and to contact protein components of the U1 and U2 snRNPs via interactions of the RS domains (Lavigne, La Branche, Kornblihtt, & Chabot, 1993; Wu & Maniatis, 1993) (Figure 1.3 b). Although SR proteins have typically been thought of as binding to exonic sequences to enhance splicing, they have also been implicated in binding to intronic sequences and in inhibiting splicing (Kanopka, Mühlemann, & Akusjärvi, 1996). Further, different SR proteins within the same organism are known to bind different classes of mRNA transcripts, so different SR proteins have specific effects during splicing (Änkö, Morales, Henry, Beyer, & Neugebauer, 2010). Another class of RNA binding proteins that are known to modulate splice site choice is the hnRNPs. In

contrast to the SR proteins, the hnRNPs are thought to largely have inhibitory effects on the usage of splice sites near where they bind (Buratti et al., 2004; Romano et al., 2002) (Figure 1.3 b), though as with SR proteins there are exceptions to this canonical mode of action (Garneau, Revil, Fiset, & Chabot, 2005; Hastings, Wilson, & Munroe, 2001). The sequences within the transcript to which the SR and hnRNP proteins bind are part of an additional class of auxiliary *cis*-acting factors that affect splicing. These sequence elements are typically short (under 10nt) and have been named and classified according to their location and the effect they have on nearby splice site usage (Black, 2003). For example, many of the sequence elements to which SR proteins bind are called exonic splicing enhancers (ESEs), since they are typically found within exons, and typically act to enhance the usage of the splice sites flanking the exon in question, via the action of the bound SR protein (Blencowe, 2000; Kohtz et al., 1994). Similarly, there are intronic splicing inhibitors (ISIs), exonic splicing inhibitors (ESIs) and intronic splicing enhancers (ISEs) (Cartegni, Chew, & Krainer, 2002; Pagani & Baralle, 2004). Collectively these auxiliary *cis* sequence elements are called splicing regulatory elements (SREs). Not all SREs are necessarily the binding targets for RNA-binding proteins; indeed many have been identified using functional screens for modulation of nearby splice site usage (Coulter, Landree, & Cooper, 1997; Wang et al., 2004), and even using

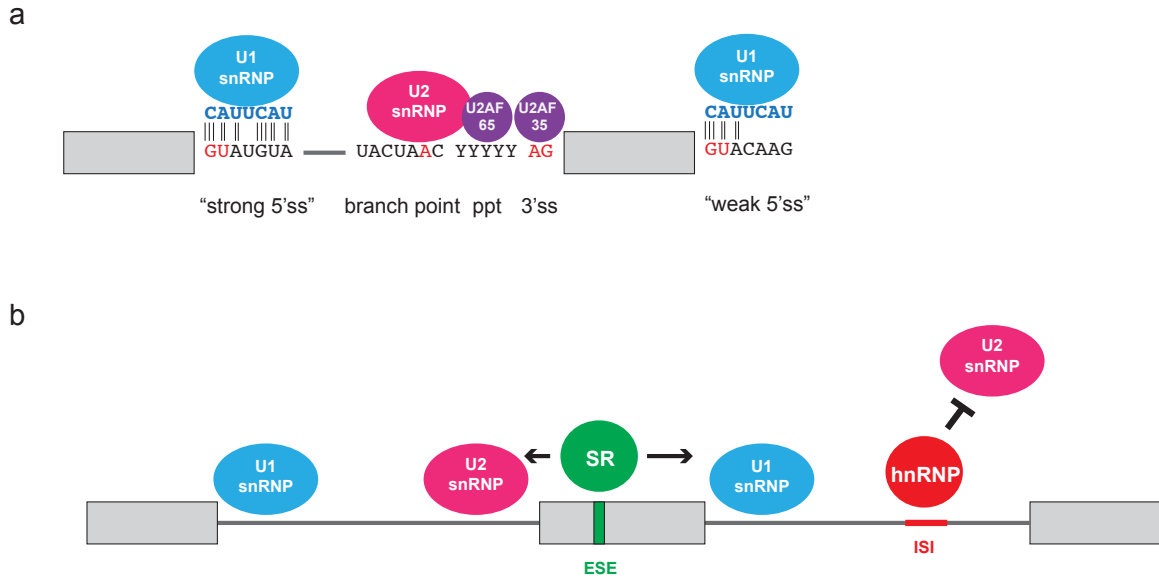


Figure 1.3

Initial assembly of the splicing machinery on introns (A) Intron 1 shown with canonical splice site sequences: five prime splice site (5'ss), branch point and three prime splice site (3'ss). Invariant residues are highlighted in red. Above intron, major components of the splicing machinery implicated in initial assembly are depicted: U1 small nuclear ribonuclear protein (snRNP) is shown with the RNA component complementary to the 5'ss shown in blue. U2 associated factor 65 (U2AF65) is shown interacting with the polypyrimidine tract (ppt) and U2AF35 with the 3'ss. A "weak" 5'ss is shown downstream, where there is diminished base pairing between the 5'ss and the RNA component of U1snRNP. (B) Auxiliary trans acting splicing factors. An SR protein is depicted binding to an exonic splicing enhancer (ESE) sequence to help recruit U2snRNP to the upstream 3'ss and U1snRNP to the downstream 5'ss. An hnRNP protein is shown binding to an intronic splicing inhibitor (ISI) sequence to prevent interaction of the U2 snRNP near the 3'ss.

computational approaches (Fairbrother, Yeh, Sharp, & Burge, 2002; Zhang & Chasin, 2004).

It has been demonstrated mathematically that if all of the potential 5' splice sites and 3' splice sites on an average transcript are considered, there is insufficient information content in the splice site sequences alone to distinguish “real” splice sites (those with experimental support) from false ones (Lim & Burge, 2001). However, even the additional information from the auxiliary cis-acting SRE sequences may be insufficient to allow prediction of splice site usage. Given the fact that these sequence elements are so short, it is unsurprising that they are highly abundant in actual intron and exon sequences, to the extent that it is difficult to predict what the overall effect of all such elements is on splicing (Buratti, Baralle, & Baralle, 2006; Chasin, 2008). Thus the spliceosome likely uses additional cues to choose between different competing potential splice sites. Two important sources of such cues are RNA structure (Jin, Yang, & Zhang, 2011; Pervouchine et al., 2012) and epigenetic information (Fox-Walsh & Fu, 2010; Luco et al., 2011). The main paradigms for RNA structure affecting splicing involve occlusion or exposure of splice sites or SREs (Singh, Singh, & Androphy, 2007; Smith & Valcárcel, 2000) and of allowing long-range RNA interactions by looping out stretches of RNA in a stem loop structure (Graveley, 2005; Miriami, Margalit, & Sperling, 2003).

As for epigenetic cues, certain histone modifications have been implicated in alternative splicing regulation by modulating recruitment of alternative splicing proteins (Luco et al., 2010), and recently DNA methylation was shown to affect the inclusion of an alternative exon (Shukla et al., 2011). The mechanism by which this was shown to occur is related to another non-genetic paradigm affecting splice-site choice: the speed of RNA polymerase II during transcription (Dujardin et al., 2012; de la Mata et al., 2003). This kinetic “first come first served” model of splice site choice posits that since splicing typically begins co-transcriptionally (Baurén & Wieslander, 1994; Tennyson, Klamut, & Worton, 1995), if a 5’ss is followed by two competing 3’ss, where the upstream 3’ss is weak and the downstream one is strong, then the choice of which 3’ss to pair to the 5’ss is affected by how quickly RNA polymerase II transcribes through that region of the gene. If RNA polymerase II transcribes quickly through that region of the gene, then the strong downstream 3’ss will be available to pair to the 5ss before the weaker upstream 3’ss can be used, whereas if transcription is slow, then there is sufficient time for the weaker upstream 3’ss to be paired to the 5’ss before the strong 3’ss is made available (de la Mata, Lafaille, & Kornblihtt, 2010).

One of the major goals of the alternative splicing field has been to use information from all of these cues to predict what the particular splicing pattern will be used for a given transcript – an endeavour that has been likened

to “deciphering the splicing code” (Matlin, Clark, & Smith, 2005; Wang & Burge, 2008). In recent years, progress has been made towards this goal (Barash et al., 2010) as well as a more modest goal: elucidating the splicing patterns for transcripts that are targeted by specific auxiliary splicing factors, given knowledge of the sequences to which these proteins bind, and the positions where these sequences occur within the transcript (Witten & Ule, 2011). In this way, “splicing maps” have been constructed for two mammalian splicing factors: NOVA and FOXO (Ule et al., 2006; Yeo et al., 2009). However, the field is still far from the goal of fully understanding how the splicing machinery decides where and when to splice. Much has been learned on this subject by studying alternative splicing in multi-cellular model eukaryotes with varying degrees of genetic tractability. On the other hand, the elucidation of the basic mechanisms of splicing and the ordering of the spliceosomal cycle benefitted enormously from the use of genetics in the highly genetically tractable unicellular eukaryote *S. cerevisiae* (Ruby & Abelson, 1991; Rymond & Rosbash, 1992). Specifically, genetic screens were used to identify temperature sensitive mutants of essential splicing factors (Bromley, Hereford, & Rosbash, 1982; Hartwell, 1967; Rosbash, Harris, Woolford Jr, & Teem, 1981), which were then used to order the spliceosomal cycle using cell free systems and biochemical assays (Cheng & Abelson, 1987; Lustig, Lin, & Abelson, 1986). This approach allowed researchers to ask and answer questions in a way that was not possible

in multi-cellular eukaryotic model organisms. Such an approach applied to the aforementioned questions of alternative splicing would likely yield important results.

1.3 Genome-wide Approaches to Study Alternative Splicing in a Genetically Tractable System

While genetic and biochemical studies in the budding yeast *Saccharomyces cerevisiae* were used to order the steps of the spliceosomal cycle as mentioned above, this organism would make a poor choice as a model system for alternative splicing in multi-cellular eukaryotes. This is because out of its ~300 intron containing genes, less than ten contain multiple introns (Cherry et al., 2012; Goffeau et al., 1996), which is a pre-requisite for the exon-skipping form of alternative splicing, and none of these genes have been shown to undergo exon skipping. Further, the splice sites in *S. cerevisiae* adhere to strict consensus sequences, with very little variation compared to multi-cellular eukaryotes such as humans (Ast, 2004; Spingola, Grate, Haussler, & Ares, 1999) (Figure 1.4). This means that, in comparison with most multi-cellular eukaryotes, there is far less potential for regulated differential splice site choice – a hallmark of alternative splicing – in *S. cerevisiae*. By contrast, another unicellular yeast, *Schizosaccharomyces pombe*, is closer phenotypically in these two traits to humans than to *S. cerevisiae*. Specifically, *S. pombe* has multiple introns in over one thousand of its genes and has splice site consensus sequences that have much more variation than those of *S. cerevisiae* (Wood et al., 2002, 2012) (Figure 1.4).

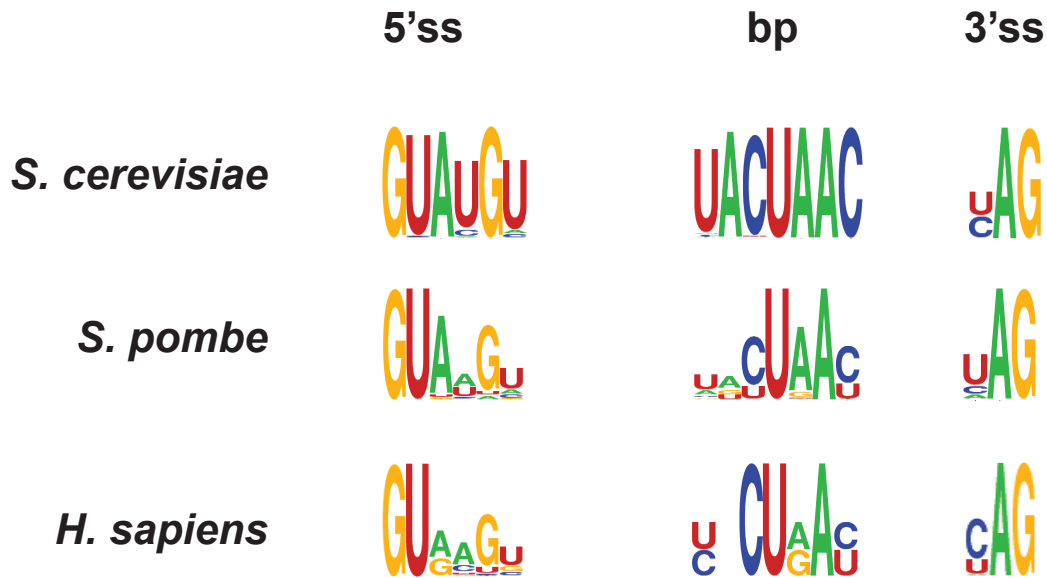


Figure 1.4

Consensus sequence logos for the main splicing signals. In each of the three species indicated on the left, all annotated introns sequences were analysed to produce consensus sequence motifs for each of three splicing sequences. The five prime splice site (5'ss) was defined as the first six nucleotides of the intron, the three prime splice site (3'ss) was defined as the last three nucleotides of the intron, and the branch point sequence is defined as seven nucleotides that encompass the inferred or experimentally determined branch point adenosine. For each of these three splice signals, consensus motifs logos were created in each species by extracting the relevant sequence from each intron and performing a multiple sequence alignment using the software weblogo (Crooks, Hon, Chandonia, & Brenner, 2004).

Moreover, *S. pombe* has a similar genetic and experimental tractability as *S. cerevisiae*, as evidenced for example by the determination of the order of the cell cycle in this organism (Nasmyth & Nurse, 1981; Nurse, Thuriaux, & Nasmyth, 1976). Further, members of the SR family of proteins that are involved in alternative splicing in multi-cellular eukaryotes have been found in *S. pombe* (Gross, Richert, Mierke, Lützelberger, & Käufer, 1998; Lützelberger, Groß, & Käufer, 1999; Tang, Käufer, & Lin, 2002) but not in *S. cerevisiae*. Finally, mammalian exonic splicing enhancer sequences (ESEs) have been shown to enhance splicing in *S. pombe* (Webb, Romfo, van Heeckeren, & Wise, 2005). These factors combine to make *S. pombe* an attractive model system in which to study alternative splicing at a genomic scale.

The most recent information about alternative splicing in *S. pombe* has been provided by genome-wide RNA-seq experiments that profiled the *S. pombe* transcriptome with hundreds of millions of short sequencing reads (Rhind et al., 2011; Wilhelm et al., 2008). While these studies revealed cases of intron retention, they found no evidence for the exon skipping form of alternative splicing. However, these results might not reflect the biology of the organism but rather technical short-comings of RNA-seq for detecting exon skipping. To address this issue, I sought to develop a different approach where sequencing power was focussed on the material excised by the splicing machinery, rather than the processed mRNA species profiled by RNA-seq.

Bibliography

- Anastasaki, C., Longman, D., Capper, A., Patton, E. E., & Cáceres, J. F. (2011). Dhx34 and nbas function in the NMD pathway and are required for embryonic development in zebrafish. *Nucleic Acids Research*, 39(9), 3686-3694.
- Ast, G. (2004). How did alternative splicing evolve? *Nature Reviews. Genetics*, 5(10), 773-82. doi:10.1038/nrg1451
- Änkö, M. -L., Morales, L., Henry, I., Beyer, A., & Neugebauer, K. M. (2010). Global analysis reveals srp20-and srp75-specific mrnps in cycling and neural cells. *Nature Structural & Molecular Biology*, 17(8), 962-970.
- Baker, B. S. (1989). Sex in flies: The splice of life. *Nature*, 340(6234), 521-524.
- Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., . . . Frey, B. J. (2010). Deciphering the splicing code. *Nature*, 465(7294), 53-59.
- Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., . . . Çolak, R. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114), 1587-1593.
- Baurén, G., & Wieslander, L. (1994). Splicing of balbiani ring 1 gene pre-mrna occurs simultaneously with transcription. *Cell*, 76(1), 183-192.

- Berget, S. M., Moore, C., & Sharp, P. A. (1977). Spliced segments at the 5'terminus of adenovirus 2 late mrna. *Proceedings of the National Academy of Sciences*, 74(8), 3171-3175.
- Bessonov, S., Anokhina, M., Will, C. L., Urlaub, H., & Lührmann, R. (2008). Isolation of an active step I spliceosome and composition of its RNP core. *Nature*, 452(7189), 846-850. Retrieved from Google Scholar.
- Bi, J., Xia, H., Li, F., Zhang, X., & Li, Y. (2005). The effect of U1 snrna binding free energy on the selection of 5 splice sites. *Biochemical and Biophysical Research Communications*, 333(1), 64-69.
- Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry*, 72(1), 291-336.
- Blencowe, B. J. (2000). Exonic splicing enhancers: Mechanism of action, diversity and role in human genetic diseases. *Trends in Biochemical Sciences*, 25(3), 106-110.
- Blencowe, B. J. (2006). Alternative splicing: New insights from global analyses. *Cell*, 126(1), 37-47. doi:10.1016/j.cell.2006.06.023
- Boise, L. H., González-Garcia, M., Postema, C. E., Ding, L., Lindsten, T., Turka, L. A., . . . Thompson, C. B. (1993). < I> bcl-x, a< i> bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell*, 74(4), 597-608.

- Bromley, S., Hereford, L., & Rosbash, M. (1982). Further evidence that the rna2 mutation of *saccharomyces cerevisiae* affects mrna processing. *Molecular and Cellular Biology*, 2(10), 1205-1211.
- Buratti, E., Baralle, M., & Baralle, F. E. (2006). Defective splicing, disease and therapy: Searching for master checkpoints in exon definition. *Nucleic Acids Research*, 34(12), 3494-510. doi:10.1093/nar/gkl498
- Buratti, E., Baralle, M., De Conti, L., Baralle, D., Romano, M., Ayala, Y. M., & Baralle, F. E. (2004). HnRNP H binding at the 5 splice site correlates with the pathological effect of two intronic mutations in the NF-1 and tsh β genes. *Nucleic Acids Research*, 32(14), 4224-4236.
- Carbone, M. A., Applegarth, D. A., & Robinson, B. H. (2002). Intron retention and frameshift mutations result in severe pyruvate carboxylase deficiency in two male siblings. *Human Mutation*, 20(1), 48-56.
- Cartegni, L., Chew, S. L., & Krainer, A. R. (2002). Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nature Reviews Genetics*, 3(4), 285-298.
- Cech, T. R. (1983). RNA splicing: Three themes with variations. *Cell*, 34(3), 713-6.
- Chasin, L. A. (2008). Searching for splicing motifs. *Advances in Experimental Medicine and Biology*, 623, 85.

- Cheng, S. C., & Abelson, J. (1987). Spliceosome assembly in yeast. *Genes & Development*, 1(9), 1014-1027.
- Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., . . . Engel, S. R. (2012). Saccharomyces genome database: The genomics resource of budding yeast. *Nucleic Acids Research*, 40(D1), D700-D705.
- Chow, L. T., Gelinas, R. E., Broker, T. R., & Roberts, R. J. (1977). An amazing sequence arrangement at the 5 ends of adenovirus 2 messenger RNA. *Cell*, 12(1), 1-8.
- Claverie, J. -M. (2001). Gene number: What if there are only 30,000 human genes? *Science Signaling*, 291(5507), 1255.
- Clements, J. A., Mercer, F. C., Paterno, G. D., & Gillespie, L. L. (2012). Differential splicing alters subcellular localization of the alpha but not beta isoform of the MIER1 transcriptional regulator in breast cancer cells. *PloS One*, 7(2), e32499.
- Collins, F. S., Lander, E. S., Rogers, J., Waterston, R. H., & Conso, I. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), 931-945.
- Coulter, L. R., Landree, M. A., & Cooper, T. A. (1997). Identification of a new class of exonic splicing enhancers by in vivo selection. *Molecular and Cellular Biology*, 17(4), 2143-2150.

Crooks, G. E., Hon, G., Chandonia, J. -M., & Brenner, S. E. (2004). WebLogo: A sequence logo generator. *Genome Research*, 14(6), 1188-1190. Retrieved from Google Scholar.

David, C. J., & Manley, J. L. (2010). Alternative pre-mrna splicing regulation in cancer: Pathways and programs unhinged. *Genes & Development*, 24(21), 2343-2364.

Dujardin, G., Lafaille, C., Petrillo, E., Buggiano, V., Gómez Acuña, L. I., Fiszbein, A., . . . Kornblihtt, A. R. (2012). Transcriptional elongation and alternative splicing. *Biochimica Et Biophysica Acta*.
doi:10.1016/j.bbagr.2012.08.005

Fairbrother, W. G., Yeh, R. -F., Sharp, P. A., & Burge, C. B. (2002). Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583), 1007-1013.

Fox-Walsh, K., & Fu, X. -D. (2010). Chromatin: The final frontier in splicing regulation? *Developmental Cell*, 18(3), 336-338.

Franzin, C. M., Yu, J., Thai, K., Choi, J., & Marassi, F. M. (2005). Correlation of gene and protein structures in the FXYD family proteins. *Journal of Molecular Biology*, 354(4), 743-750.

Garneau, D., Revil, T., Fiset, J. -F., & Chabot, B. (2005). Heterogeneous nuclear ribonucleoprotein F/H proteins modulate the alternative splicing of the apoptotic mediator bcl-x. *Journal of Biological Chemistry*, 280(24), 22641-22650.

- Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., . . . Johnston, M. (1996). Life with 6000 genes. *Science*, 274(5287), 546-567.
- Goulet, I., Gauvin, G., Boisvenue, S., & Côté, J. (2007). Alternative splicing yields protein arginine methyltransferase 1 isoforms with distinct activity, substrate specificity, and subcellular localization. *Journal of Biological Chemistry*, 282(45), 33009-33021.
- Graveley, B. R. (2001). Alternative splicing: Increasing diversity in the proteomic world. *Trends in Genetics : TIG*, 17(2), 100-7.
- Graveley, B. R. (2005). Mutually exclusive splicing of the insect *dscam* pre-mrna directed by competing intronic RNA secondary structures. *Cell*, 123(1), 65-73.
- Gross, T., Richert, K., Mierke, C., Lützelberger, M., & Käufer, N. F. (1998). Identification and characterization of *srp1*, a gene of fission yeast encoding a RNA binding domain and a RS domain typical of SR splicing factors. *Nucleic Acids Research*, 26(2), 505-511.
- Guthrie, C. (1994). The spliceosome is a dynamic ribonucleoprotein machine. *Harvey Lectures*, 90, 59-80.
- Hartwell, L. H. (1967). Macromolecule synthesis in temperature-sensitive mutants of yeast. *Journal of Bacteriology*, 93(5), 1662-1670.

Hastings, M. L., Wilson, C. M., & Munroe, S. H. (2001). A purine-rich intronic element enhances alternative splicing of thyroid hormone receptor mrna. *RNA (New York, N.Y.)*, 7(6), 859-874.

Hattori, D., Demir, E., Kim, H. W., Viragh, E., Zipursky, S. L., & Dickson, B. J. (2007). Dscam diversity is essential for neuronal wiring and self-recognition. *Nature*, 449(7159), 223-227.

Hattori, D., Millard, S. S., Wojtowicz, W. M., & Zipursky, S. L. (2008). Dscam-mediated cell recognition regulates neural circuit formation. *Annual Review of Cell and Developmental Biology*, 24, 597.

Hayashi, T., Gekka, T., Omoto, S., Takeuchi, T., & Kitahara, K. (2005).

Dominant optic atrophy caused by a novel OPA1 splice site mutation (IVS20+1G A) associated with intron retention. *Ophthalmic Research*, 37(4), 214-224.

doi:10.1159/000086862

Hentze, M. W., & Kulozik, A. E. (1999). A perfect message: RNA surveillance and nonsense-mediated decay. *Cell*, 96(3), 307-310.

Jin, Y., Yang, Y., & Zhang, P. (2011). New insights into RNA secondary structure in the alternative splicing of pre-mrnas. *RNA Biology*, 8(3), 450-457.

Jurica, M. S., Licklider, L. J., Gygi, S. R., Grigorieff, N., & Moore, M. J. (2002).

Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. *RNA (New York, N.Y.)*, 8(4), 426. Retrieved from Google Scholar.

Kanopka, A., Mühlemann, O., & Akusjärvi, G. (1996). Inhibition by SR proteins of splicing of a regulated adenovirus pre-mrna. *Nature*, 381(6582), 535-538.

Kohtz, J. D., Jamison, S. F., Will, C. L., Zuo, P., Lührmann, R., Garcia-Blanco, M. A., & Manley, J. L. (1994). Protein--protein interactions and 5'-splice-site recognition in mammalian mrna precursors. *Protein--protein Interactions and 5'-splice-site Recognition in Mammalian MRNA Precursors*.

Konarska, M. M., & Query, C. C. (2005). Insights into the mechanisms of splicing: More lessons from the ribosome. *Genes & Development*, 19(19), 2255-60. doi:10.1101/gad.1363105

de la Mata, M., Alonso, C. R., Kadener, S., Fededa, J. P., Blaustein, M., Pelisch, F., . . . Kornblihtt, A. R. (2003). A slow RNA polymerase II affects alternative splicing in vivo. *Molecular Cell*, 12(2), 525-532.

de la Mata, M., Lafaille, C., & Kornblihtt, A. R. (2010). First come, first served revisited: Factors affecting the same alternative splicing event have different effects on the relative rates of intron removal. *RNA (New York, N.Y.)*, 16(5), 904-912.

Lavigne, A., La Branche, H., Kornblihtt, A. R., & Chabot, B. (1993). A splicing enhancer in the human fibronectin alternate ED1 exon interacts with SR proteins and stimulates U2 snrnp binding. *Genes & Development*, 7(12a), 2405-2417.

- Lim, L. P., & Burge, C. B. (2001). A computational analysis of sequence features involved in recognition of short introns. *Proceedings of the National Academy of Sciences*, 98(20), 11193-11198.
- Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R., & Misteli, T. (2011). Epigenetics in alternative pre-mrna splicing. *Cell*, 144(1), 16-26.
doi:10.1016/j.cell.2010.11.056
- Luco, R. F., Pan, Q., Tominaga, K., Blencowe, B. J., Pereira-Smith, O. M., & Misteli, T. (2010). Regulation of alternative splicing by histone modifications. *Science (New York, N.Y.)*, 327(5968), 996-1000. doi:10.1126/science.1184208
- Lustig, A. J., Lin, R. -J., & Abelson, J. (1986). The yeast *RNA* gene products are essential for mrna splicing in vitro. *Cell*, 47(6), 953-963.
- Lützelberger, M., Groß, T., & Käufer, N. F. (1999). Srp2, an SR protein family member of fission yeast: In vivo characterization of its modular domains. *Nucleic Acids Research*, 27(13), 2618-2626.
- Madhani, H. D., & Guthrie, C. (1994). Dynamic RNA-RNA interactions in the spliceosome. *Annual Review of Genetics*, 28, 1-26.
doi:10.1146/annurev.ge.28.120194.000245
- Maniatis, T., & Tasic, B. (2002). Alternative pre-mrna splicing and proteome expansion in metazoans. *Nature*, 418(6894), 236-43. doi:10.1038/418236a

- Maquat, L. E. (1995). When cells stop making sense: Effects of nonsense codons on RNA metabolism in vertebrate cells. *RNA (New York, N.Y.)*, 1(5), 453-465.
- Matlin, A. J., Clark, F., & Smith, C. W. (2005). Understanding alternative splicing: Towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6(5), 386-398.
- Michael, I. P., Kurlender, L., Memari, N., Yousef, G. M., Du, D., Grass, L., . . . Diamandis, E. P. (2005). Intron retention: A common splicing event within the human kallikrein gene family. *Clinical Chemistry*, 51(3), 506-515.
- Miriami, E., Margalit, H., & Sperling, R. (2003). Conserved sequence elements associated with exon skipping. *Nucleic Acids Research*, 31(7), 1974-1983.
- Nasmyth, K., & Nurse, P. (1981). Cell division cycle mutants altered in DNA replication and mitosis in the fission yeast *schizosaccharomyces pombe*. *Molecular and General Genetics MGG*, 182(1), 119-124.
- Nilsen, T. W. (1998). RNA-RNA interactions in nuclear pre-mrna splicing. *Cold Spring Harbor Monograph Archive*, 35, 279-307. Retrieved from Google Scholar.
- Nilsen, T. W., & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280), 457-63. doi:10.1038/nature08909
- Nott, A., Le Hir, H., & Moore, M. J. (2004). Splicing enhances translation in mammalian cells: An additional function of the exon junction complex. *Genes & Development*, 18(2), 210-222.

- NOTT, A., MEISLIN, S. H., & MOORE, M. J. (2003). A quantitative analysis of intron effects on mammalian gene expression. *RNA (New York, N.Y.)*, 9(5), 607-617.
- Nurse, P., Thuriaux, P., & Nasmyth, K. (1976). Genetic control of the cell division cycle in the fission yeast *schizosaccharomyces pombe*. *Molecular and General Genetics MGG*, 146(2), 167-178.
- Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S., & Sharp, P. A. (1986). Splicing of messenger RNA precursors. *Annual Review of Biochemistry*, 55, 1119-50. doi:10.1146/annurev.bi.55.070186.005351
- Pagani, F., & Baralle, F. E. (2004). Genomic variants in exons and introns: Identifying the splicing spoilers. *Nature Reviews Genetics*, 5(5), 389-396.
- Papadopoulos, C., Arato, K., Lilienthal, E., Zerweck, J., Schutkowski, M., Chatain, N., . . . de la Luna, S. (2011). Splice variants of the dual specificity tyrosine phosphorylation-regulated kinase 4 (DYRK4) differ in their subcellular localization and catalytic activity. *Journal of Biological Chemistry*, 286(7), 5494-5505.
- Pervouchine, D. D., Khrameeva, E. E., Pichugina, M. Y., Nikolaienko, O. V., Gelfand, M. S., Rubtsov, P. M., & Mironov, A. A. (2012). Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. *RNA (New York, N.Y.)*, 18(1), 1-15.

- Pleiss, J. A., Whitworth, G. B., Bergkessel, M., & Guthrie, C. (2007). Rapid, transcript-specific changes in splicing in response to environmental stress. *Molecular Cell*, 27(6), 928-37. doi:10.1016/j.molcel.2007.07.018
- Pollard, A. J., Flanagan, B. F., Newton, D. J., & Johnson, P. M. (1998). A novel isoform of human membrane cofactor protein (CD46) mrna generated by intron retention. *Gene*, 212(1), 39-47.
- Puig, O., Gottschalk, A., Fabrizio, P., & Séraphin, B. (1999). Interaction of the U1 snrnp with nonconserved intronic sequences affects 5 splice site selection. *Genes & Development*, 13(5), 569-580.
- Rhind, N., Chen, Z., Yassour, M., Thompson, D. A., Haas, B. J., Habib, N., . . . Heiman, D. I. (2011). Comparative functional genomics of the fission yeasts. *Science*, 332(6032), 930. doi:10.1126/science.1203357
- Riccombeni, A., Vidanes, G., Proux-Wéra, E., Wolfe, K. H., & Butler, G. (2012). Sequence and analysis of the genome of the pathogenic yeast candida orthopsilosis. *PloS One*, 7(4), e35750.
- ROCA, X., SACHIDANANDAM, R., & KRAINER, A. R. (2005). Determinants of the inherent strength of human 5 splice sites. *RNA (New York, N.Y.)*, 11(5), 683-698.
- Romano, M., Marcucci, R., Buratti, E., Ayala, Y. M., Sebastio, G., & Baralle, F. E. (2002). Regulation of 3 splice site selection in the 844ins68 polymorphism of

the cystathionine β -synthase gene. *Journal of Biological Chemistry*, 277(46), 43821-43829.

De Roos, A. D. (2007). Conserved intron positions in ancient protein modules. *Biology Direct*, 2(7).

Rosbash, M., Harris, P. K., Woolford Jr, J. L., & Teem, J. L. (1981). The effect of temperature-sensitive RNA mutants on the transcription products from cloned ribosomal protein genes of yeast. *Cell*, 24(3), 679-686.

Ruby, S. W., & Abelson, J. (1991). Pre-mRNA splicing in yeast. *Trends in Genetics*, 7(3), 79-85.

Rymond, B. C., & Rosbash, M. (1992). 4 yeast pre-mrna splicing. *Cold Spring Harbor Monograph Archive*, 21, 143-192.

Schulze-Osthoff, K., Ferrari, D., Los, M., Wesselborg, S., & Peter, M. E. (1998). Apoptosis signaling by death receptors. *European Journal of Biochemistry*, 254(3), 439-459.

Schutt, C., & Nothiger, R. (2000). Structure, function and evolution of sex-determining systems in dipteran insects. *Development*, 127(4), 667-677.

Senapathy, P. (1986). Origin of eukaryotic introns: A hypothesis, based on codon distribution statistics in genes, and its implications. *Proceedings of the National Academy of Sciences*, 83(7), 2133-2137.

Shepard, P. J., & Hertel, K. J. (2009). The SR protein family. *Genome Biology*, 10(10), 242.

- Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., . . . Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, *479*(7371), 74-9.
doi:10.1038/nature10442
- Singh, N. N., Singh, R. N., & Androphy, E. J. (2007). Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. *Nucleic Acids Research*, *35*(2), 371-389.
- Singh, R. K., & Cooper, T. A. (2012). Pre-mRNA splicing in disease and therapeutics. *Trends in Molecular Medicine*.
- Smith, C. W., & Valcárcel, J. (2000). Alternative pre-mrna splicing: The logic of combinatorial control. *Trends in Biochemical Sciences*, *25*(8), 381-387.
- Spingola, M., Grate, L., Haussler, D., & Ares, M. (1999). Genome-wide bioinformatic and molecular analysis of introns in *saccharomyces cerevisiae*. *RNA (New York, N.Y.)*, *5*(2), 221-234.
- Spitali, P., & Aartsma-Rus, A. (2012). Splice modulating therapies for human disease. *Cell*, *148*(6), 1085-1088.
- Staley, J. P., & Guthrie, C. (1998). Mechanical devices of the spliceosome: Review motors, clocks, springs, and things. *Cell*, *92*, 315-326. Retrieved from Google Scholar.
- Stein, L. D. (2004). Human genome: End of the beginning. *Nature*, *431*(7011), 915-916.

- Strijbis, K., den Burg, J., F Visser, W., den Berg, M., & Distel, B. (2012). Alternative splicing directs dual localization of candida albicans 6-phosphogluconate dehydrogenase to cytosol and peroxisomes. *FEMS Yeast Research*, 12(1), 61-68.
- Tang, Z., Käufer, N. F., & Lin, R. -J. (2002). Interactions between two fission yeast serine/arginine-rich proteins and their modulation by phosphorylation. *Biochemical Journal*, 368(Pt 2), 527.
- Tennyson, C. N., Klamut, H. J., & Worton, R. G. (1995). The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nature Genetics*, 9(2), 184-190.
- Terenzi, F., & Ladd, A. N. (2010). Conserved developmental alternative splicing of muscleblind-like (MBNL) transcripts regulates MBNL localization and activity. *RNA Biology*, 7(1), 43-55.
- Torahiko, N., Takeshi, S., Hidetoshi, S., Kei, Y., Shirou, T., Mitiko, G., . . . Takeharu, N. (1994). Structure of the human CCG1 gene: Relationship between the exons/introns and functional domain/modules of the protein. *Gene*, 141(2), 193-200.
- Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., . . . Darnell, R. B. (2006). An RNA map predicting nova-dependent splicing regulation. *Nature*, 444(7119), 580-586.

- Wahl, M. C., Will, C. L., & Lührmann, R. (2009). The spliceosome: Design principles of a dynamic RNP machine. *Cell*, *136*(4), 701-18.
doi:10.1016/j.cell.2009.02.009
- Wang, E. T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., . . . Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470-476.
- Wang, Z., & Burge, C. B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA (New York, N.Y.)*, *14*(5), 802-813.
- Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M., & Burge, C. B. (2004). Systematic identification and analysis of exonic splicing silencers. *Cell*, *119*(6), 831-845.
- Webb, C. J., Romfo, C. M., van Heeckeren, W. J., & Wise, J. A. (2005). Exonic splicing enhancers in fission yeast: Functional conservation demonstrates an early evolutionary origin. *Genes & Development*, *19*(2), 242-54.
doi:10.1101/gad.1265905
- Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., . . . Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, *453*(7199), 1239-1243.
- Witten, J. T., & Ule, J. (2011). Understanding splicing regulation through RNA splicing maps. *Trends in Genetics*, *27*(3), 89-97.

Wittkopp, N., Huntzinger, E., Weiler, C., Saulière, J., Schmidt, S., Sonawane, M., & Izaurralde, E. (2009). Nonsense-mediated mrna decay effectors are essential for zebrafish embryonic development and survival. *Molecular and Cellular Biology*, 29(13), 3517-3528.

Wood, V., Gwilliam, R., Rajandream, M. -A., Lyne, M., Lyne, R., Stewart, A., . . . Baker, S. (2002). The genome sequence of schizosaccharomyces pombe. *Nature*, 415(6874), 871-880.

Wood, V., Harris, M. A., McDowall, M. D., Rutherford, K., Vaughan, B. W., Staines, D. M., . . . Kersey, P. J. (2012). PomBase: A comprehensive online resource for fission yeast. *Nucleic Acids Research*, 40(D1), D695-D699.

Wu, J. Y., & Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell*, 75(6), 1061-1070.

Yeo, G. W., Coufal, N. G., Liang, T. Y., Peng, G. E., Fu, X. -D., & Gage, F. H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping rna-protein interactions in stem cells. *Nature Structural & Molecular Biology*, 16(2), 130-137.

Zahler, A. M., Lane, W. S., Stolk, J. A., & Roth, M. B. (1992). SR proteins: A conserved family of pre-mrna splicing factors. *Genes & Development*, 6(5), 837-847.

Zhang, X. H., & Chasin, L. A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. *Genes & Development*, 18(11), 1241-1250.

CHAPTER 2: DEVELOPING LARIAT SEQUENCING TO DETECT EXON SKIPPING IN *S. POMBE*

2.1 Introduction

The protein coding regions of eukaryotic genes are typically interrupted by non-coding introns that must be removed to produce a translatable mRNA. The removal of introns, catalyzed by the spliceosome, offers a powerful opportunity for an organism to regulate gene expression. In mammals, where individual genes are often interrupted by multiple introns, it is now abundantly clear that the process of intron removal provides a critical regulatory control point for both qualitative and quantitative aspects of gene expression (1). By changing the identity of the exons that are included within the final mRNA, the process of alternative splicing plays a critical role in expanding the diversity of proteins that can be synthesized within a cell (2). Moreover, alternative splicing can direct the production of isoforms of genes that are directly targeted to cellular decay pathways, providing a mechanism to quantitatively regulate gene expression (3, 4).

In mammalian organisms the predominant form of alternative splicing is exon skipping, wherein different combinations of exons are included in the final transcript. By contrast, exon skipping is far less prevalent in simpler

eukaryotes (5, 6); however recent studies suggest that splicing in the last eukaryotic common ancestor was similar in many respects to splicing in vertebrates, in so much as it was intron-dense (7–9), had degenerate splice site sequences (10), and likely had many of the proteins involved in alternative splicing (11, 12). Intron density correlates positively with the prevalence of alternative splicing across the eukaryotic kingdoms (13) and thus it has been posited that the intron-rich eukaryotic ancestor had alternative splicing, and may have had exon skipping. However, it is far less clear whether the inferred ancestral exon skipping was functional or represented “spliceosomal noise” that was later harnessed in the evolutionary lineages that led to multicellularity (13, 14). Two “smoking guns” that would corroborate early functional exon skipping have been previously suggested: environmentally- or developmentally-regulated exon skipping in a unicellular eukaryote; and conservation of a particular exon skipping pattern across eukaryotic kingdoms (11, 14). While examples of exon skipping in a few unicellular eukaryotes have been published (15–17), to our knowledge there are no published examples of either evolutionarily conserved or environmentally regulated exon skipping events in any unicellular eukaryote.

The unicellular fission yeast, *Schizosaccharomyces pombe*, shares many of the hallmarks of alternative splicing in mammalian systems. Nearly 50% of *S. pombe* genes contain an intron, and almost half of those contain multiple introns (18,

19). Moreover, the splice site sequences found within *S. pombe* introns do not conform to the tight consensus sequences seen in some other unicellular fungi, but rather are marked by a degeneracy more similar to that seen in human introns (20, 21). Importantly, a single bona fide member of the SR family of splicing regulators and several SR-like proteins implicated in the regulation of alternative splicing are encoded within the *S. pombe* genome (22, 23). Further, cis-acting sequence elements important in splicing regulation in multicellular eukaryotes have been shown to modulate the splicing efficiency of individual *S. pombe* introns (24), and indeed non-endogenous plant and mammalian intron sequences have been shown to be properly excised in *S. pombe* (25, 26). Nevertheless, while instances of intron-retention have been documented (27, 28), no examples of exon-skipping have been described in *S. pombe*. Surprisingly, two recent RNA-seq studies (29, 30) failed to detect any instances of exon-skipping in *S. pombe*; however, it was unclear whether this reflected an absence of such splicing events in *S. pombe*, or a limitation of RNA-seq for detecting them.

2.2 Results

Lariat Sequencing

To facilitate a deeper analysis of global splicing patterns in *S. pombe* that might uncover exon-skipping events, we developed an alternative approach that allows sequencing efforts to be concentrated on the products of a splicing reaction. Rather than sequencing the ligated mRNA product of a splicing reaction, this approach enriches, purifies, and sequences the excised lariat RNAs that are a product of every splicing reaction. Significant stabilization of lariat RNAs was achieved by genetically deleting the gene encoding the debranching enzyme (Dbr1) responsible for linearizing the introns (31), and circular RNAs were separated from linear RNAs using two-dimensional (2D) polyacrylamide gel electrophoresis (32). Because environmental stresses can stimulate splicing regulation even in budding yeast (33), a $\Delta dbr1$ *S. pombe* strain was exposed to a variety of environmental stresses, including various nutrient deprivations, temperature variations, and chemical exposures (Table 2.1, Methods). Untreated total RNA from each individual sample was pooled together and subjected to 2D gel electrophoresis (see Methods).

In order to increase the chances of identifying exon-skipping events, gel conditions were optimized for recovery of RNAs in a size range likely to include the majority of lariats that contained skipped exons. According to the

Table 2.1 - Conditions under which *Δdbr1 S. pombe* cultures were grown and/or perturbed prior to RNA isolation

| sample # | media | condition | temperature (°C) | Final OD | minutes | Stressor concentration |
|----------|-------|-------------------|------------------|----------|---------|------------------------|
| 1 | YES | vegetative growth | 30 | 0.1 | | |
| 2 | YES | vegetative growth | 30 | 0.6 | | |
| 3 | YES | vegetative growth | 30 | 1 | | |
| 4 | YES | vegetative growth | 30 | 5 | | |
| 5 | YES | vegetative growth | 25 | 0.6 | | |
| 6 | YES | vegetative growth | 37 | 0.6 | | |
| 7 | YES | vegetative growth | 16 | 0.6 | | |
| 8 | YES | heat shock | 25-42 | 0.6 | 10 | |
| 9 | YES | heat shock | 25-42 | 0.6 | 60 | |
| 10 | YES | cold shock | 30-16 | 0.6 | 30 | |
| 11 | YES | cold shock | 30-16 | 0.6 | 180 | |
| 12 | YES | salt stress-KCl | 30 | 0.6 | 20 | 0.5M |
| 13 | YES | salt stress-KCl | 30 | 0.6 | 120 | 0.5M |
| 14 | YES | DNA dmg-4NQO | 30 | 0.6 | 20 | 10 ng/mL |
| 15 | YES | DNA dmg-4NQO | 30 | 0.6 | 120 | 10 ng/mL |
| 16 | YES | DNA dmg-MMS | 30 | 0.6 | 20 | 0.02% |
| 17 | YES | DNA dmg-MMS | 30 | 0.6 | 120 | 0.02% |
| 18 | YES | Lithium | 30 | 0.6 | 20 | 0.1 M |
| 19 | YES | Lithium | 30 | 0.6 | 120 | 0.1 M |
| 20 | YES | Ethanol | 30 | 0.6 | 20 | 10% |
| 21 | YES | Ethanol | 30 | 0.6 | 120 | 10% |
| 22 | YES | DTT | 30 | 0.6 | 20 | 2.5 mM |
| 23 | YES | DTT | 30 | 0.6 | 120 | 2.5 mM |

| | | | | | | |
|----|----------|-------------------------------|----|-----|--------|--------|
| 24 | YES | H ₂ O ₂ | 30 | 0.6 | 20 | 0.32mM |
| 25 | YES | H ₂ O ₂ | 30 | 0.6 | 120 | 0.32mM |
| 26 | EMM | vegetative growth | 30 | 0.1 | | |
| 27 | EMM | vegetative growth | 30 | 0.6 | | |
| 28 | EMM | vegetative growth | 30 | 1 | | |
| 29 | EMM | vegetative growth | 30 | 5 | | |
| 30 | EMM | vegetative growth | 25 | 0.6 | | |
| 31 | EMM | vegetative growth | 37 | 0.6 | | |
| 32 | EMM | vegetative growth | 16 | 0.6 | | |
| 33 | EMM | 3-AT | 30 | 0.6 | 30 | 50 mM |
| 34 | EMM | 3-AT | 30 | 0.6 | 180 | 50 mM |
| 35 | EMM | no glucose | 30 | 0.6 | 20 | |
| 36 | EMM | no glucose | 30 | 0.6 | 120 | |
| 37 | EMM | no Nitrogen | 30 | 0.6 | 20 | |
| 38 | EMM | no Nitrogen | 30 | 0.6 | 120 | |
| 39 | EMM | no Phosphorus | 30 | 0.6 | 20 | |
| 40 | EMM | no Phosphorus | 30 | 0.6 | 120 | |
| 41 | YES-Agar | solid media | 30 | | 2 days | |

most recent *S. pombe* genome annotation, the shortest possible lariat that could be formed by skipping a single exon is 89 nucleotides. As such, for this work gel conditions were determined to optimize recovery of lariats ~85 nucleotides or greater (see Methods). The conditions used here successfully isolated lariats up to ~400 nucleotides in length, enabling detection of 66% of all possible single exon-skipping events in the *S. pombe* genome (Table 2.2*, Methods).

Lariat RNAs are detected in 2D gels as an arc above a diagonal of linear RNAs (Figure 2.1a). This “lariat arc” was excised from the 2D gels and converted into cDNA by random priming without prior debranching. The resulting cDNAs were sequenced using Illumina’s 3G technology, and the resulting reads were aligned to the *S. pombe* genome using Bowtie (34). Importantly, out of a total of 12.4 million alignable reads, over 80% of the reads derived from this approach align to previously annotated introns (Table 2.3), confirming the enrichment of lariat RNAs. “Peaks” of read enrichment were subsequently identified, allowing for the characterization of putative lariat RNAs. In considering only those peaks with a minimum of 25 overlapping reads, lariats corresponding to 84% of all annotated introns between 85 and 400 nucleotides were identified, confirming the efficiency of lariat purification. Representative examples of lariat sequencing read peaks aligning to known introns are shown in Figure 2.1b for the *sim4* and *rpL4301* genes, representing the lower and upper ends of read counts, respectively (Table 2.4*).

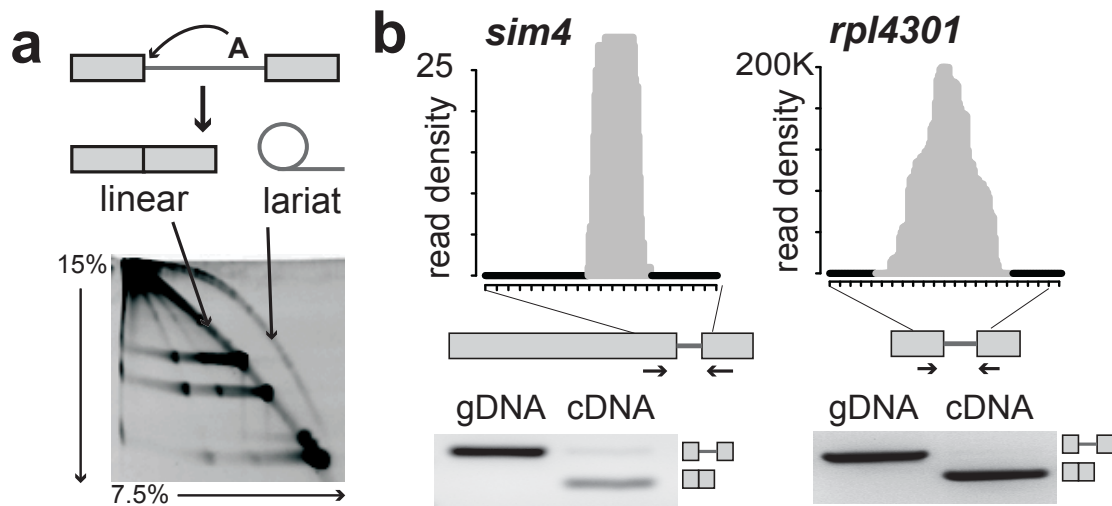


Figure 2.1

Purification and sequencing of excised lariat introns. **(a)** Stabilization of lariat RNAs is achieved by genetically deleting the gene encoding the debranching enzyme, Dbr1. Total cellular RNA isolated from a $\Delta dbr1$ strain grown under a variety of conditions was pooled and subjected to two-dimensional gel electrophoresis, allowing for separation of linear RNAs from circular lariats. **(b)** Lariat RNAs recovered from the two dimensional gel were sequenced and aligned to the *S. pombe* genome. Read density plots are shown near the known introns in the *sim4* and *rpl4301* genes. Arrows indicate locations of primers used for confirmation. PCR products generated using either genomic DNA or cDNA as a template confirm the splicing of these introns. Locations of the unspliced and spliced products are noted.

To validate splicing of these or other putative introns, flanking PCR was used wherein primers that flank predicted splice sites were used in PCR reactions containing either genomic DNA or cDNA as template. For splicing events validated using flanking PCR in this study, the genomic and cDNA templates were generated from a wild type strain containing full Dbr1 activity, grown under standard growth conditions, unless otherwise noted. As expected, the spliced isoform is the predominant species for both of these genes.

Lariat sequencing identifies over 200 novel splicing events

Interestingly, although over 80% of the lariat sequencing reads mapped to previously annotated introns, nearly 15% of the reads mapped to regions currently annotated as untranslated or even protein-coding (Table 2.3). To determine whether these peaks represented novel, unannotated splicing events, probabilistic modeling was used to score putative splicing motifs surrounding each read enrichment peak in these regions. A Markov model derived from known intron sequences (35) allowed for the quality of putative introns to be scored (see Methods), enabling identification of over 200 new introns (Table 2.5), the majority of which have splice site sequences that are indistinguishable from canonical *S. pombe* introns (Figure 2.S1).

Nearly half of the novel introns identified here are located within the untranslated regions (UTR) of protein coding genes, two of which are shown in Figure 2a. For the *cnl2* gene, the peak lies completely within the 5' UTR, ending

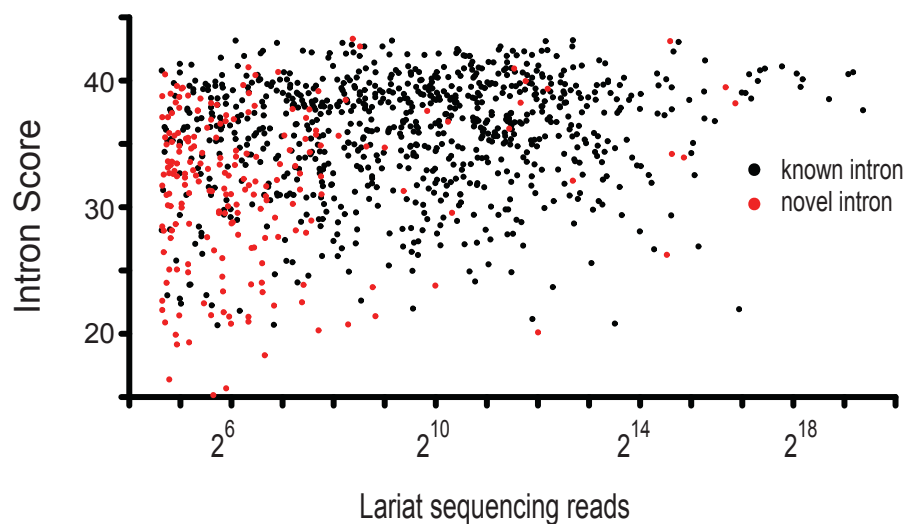


Figure 2.S1

For each known (black) and novel (red) intron with length between 85 and 200 nucleotides, a probabilistic intron score was determined (see Methods), with higher scores indicating a higher similarity to the consensus sequences. Scores are plotted relative to the lariat sequencing read count.

Table 2.3 - Alignable sequencing read counts by genomic feature

| Genomic Feature Type | Read Count |
|----------------------|------------|
| intronic | 10060519 |
| UTR or exonic | 797570 |
| rRNA | 281310 |
| other ncRNA | 1238735 |
| Total | 12378134 |

7 nucleotides upstream of the translational start codon. Similarly, for the *caf5* gene the peak lies entirely within the 5' UTR, ending 33 nucleotides upstream of the translational start codon. For both of these putative introns, strong splice site sequences were identified at the boundaries of the lariat sequencing peaks. Flanking PCR in a wild type strain confirms the splicing of both of these introns. Moreover, in the background of a strain lacking the non-essential splicing factor *smd3*, the efficiency of each of these splicing events is reduced, consistent with their being canonical, spliceosomal introns. Importantly, 53 of these novel introns are associated with genes currently annotated as intron-less, including *caf5*, expanding the number of *S. pombe* genes whose transcripts require spliceosomal processing, and which may in turn be subject to spliceosomal regulation. A subset of the predicted introns that vary both in terms of lariat sequencing read count and probabilistic modeling score was chosen for further validation, each of which confirms the splicing event, suggesting a high true positive rate of discovery (Figure 2.S2). Our data suggest that as many as 10% of all *S. pombe* genes contain an intron within their 5' or 3' UTRs. While such introns have often been ignored, recent work strongly argues in favor of a role for these introns in regulating gene expression (36).

A second category of peaks was identified in which the putative intron was located entirely within a coding region, suggesting the existence of alternatively retained introns. Twenty five such peaks were identified, two of which are

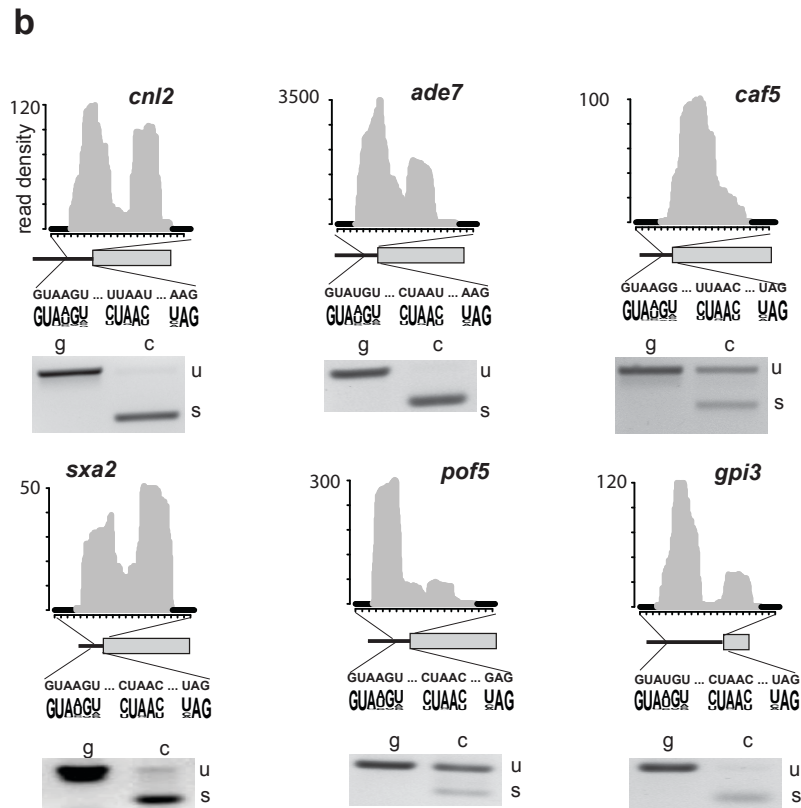
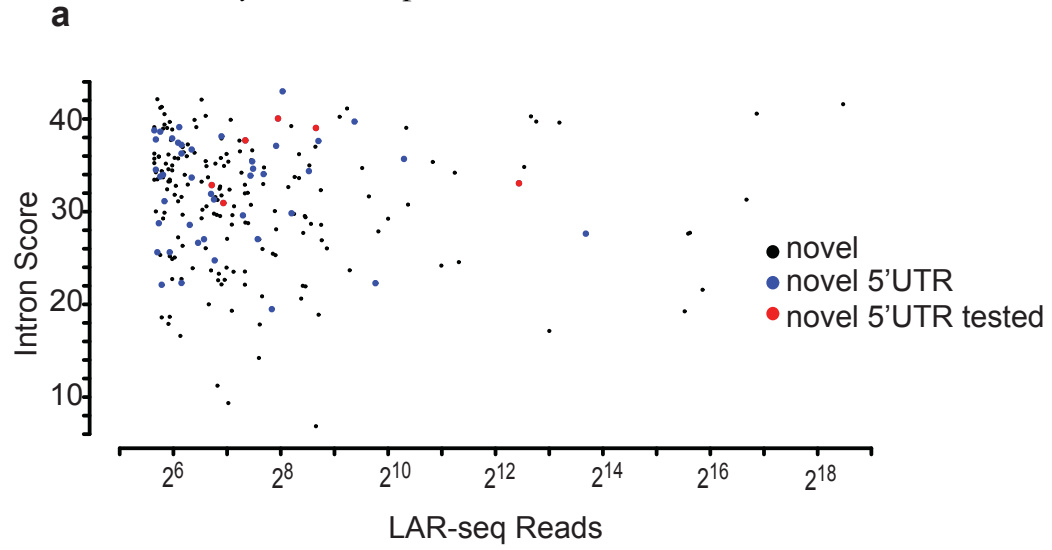


Figure 2.S2

Novel 5'UTR introns predicted by lariat sequencing are validated using flanking PCR. **(a)** Plot of probabilistic intron score versus peak read count for all novel introns between 85 and 200nt in length (black), highlighting those introns that are found within 5'UTRs (blue) and those which were tested for splicing using flanking PCR (red). **(b)** Read density plots and PCR confirmation of novel introns in the *cnl2*, *ade7*, *caf5*, *sxa2*, *pof5* and *gpi3* 5'UTRs.

shown in Figure 2.2b. For six of these peaks, with read depths between 60 and 500, the putative intron has a length that is divisible by three, as seen for the peak in the *cys11* gene, suggesting that its removal would not disrupt the translational frame of the downstream portion of the protein. Flanking PCR revealed an intermediate splicing phenotype of this intron, with both the spliced and the retained isoforms readily detectable. By contrast, 19 of the exonic peaks, with read depths between 50 and 600, suggest the presence of an intron whose length is not divisible by three, as seen for *rad8* in Figure 2.2b. Removal of these introns is predicted to change the reading frame, in all cases resulting in production of an mRNA with a premature stop codon. While such isoforms may generate truncated proteins, they are often targeted to the nonsense mediated decay (NMD) pathway (37). Flanking PCR examining the novel *rad8* intron revealed only a small amount of spliced product in a wild type strain, but an increase in a strain where the NMD factor *upf1* (38) had been disrupted (Figure 2.2b), suggesting that a sizable fraction of *rad8* transcripts are normally spliced at this intron and that splicing may be an important control point for its regulation.

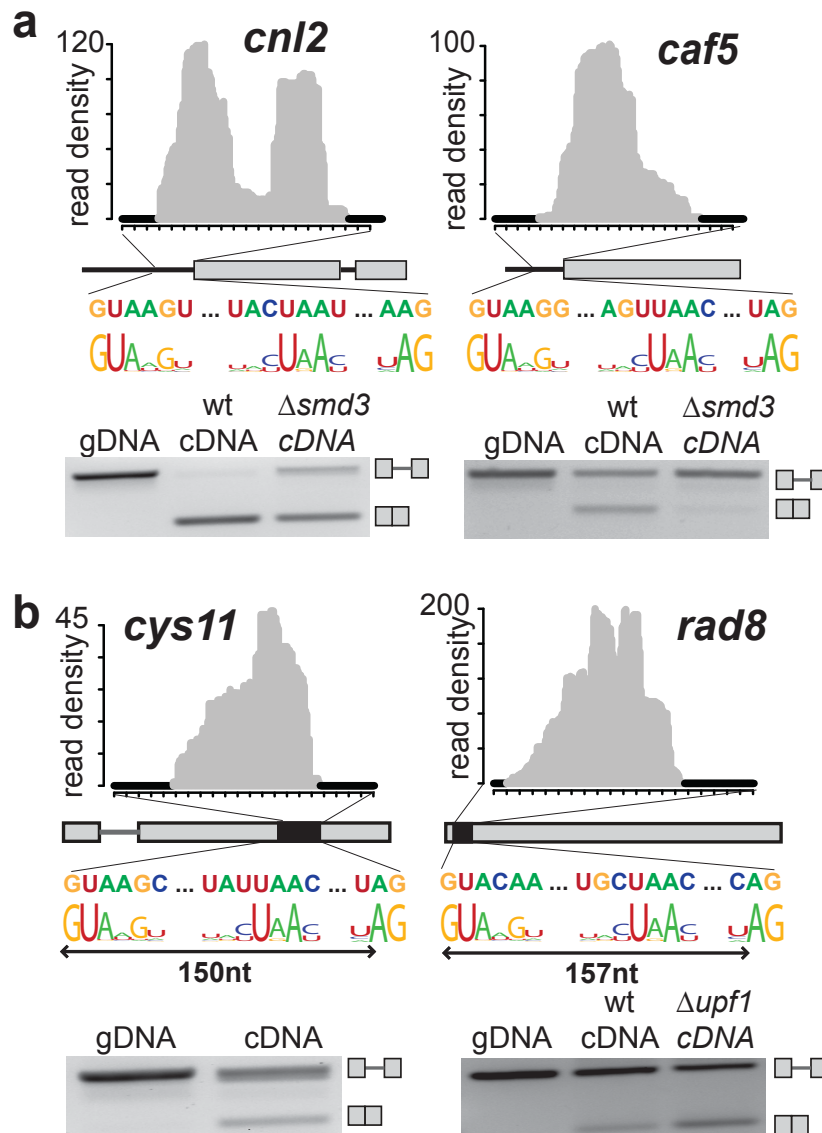


Figure 2.2

Lariat sequencing identifies over 200 novel introns. **(a)** Read density plots of new introns in the *cnl2* and *caf5* genes. Cartoons of the currently annotated coding regions are indicated below. The sequences corresponding to the putative branch point, 5' and 3' splice sites are indicated, as are the consensus sequences in known *S. pombe* introns (59). Flanking PCR demonstrates removal of the introns in cDNA from wild type cells, but a block to splicing in cDNA from cells lacking the splicing factor, *smd3* (Δ *smd3* cDNA) **(b)** Read density plots of novel introns in the coding regions of the *cys11* and *rad8* genes. Intron length is indicated by arrows. Flanking PCR using cDNA from wild type cells reveals an intermediate amount of splicing of these introns. In the background

of a strain lacking the NMD factor *upf1* ($\Delta upf1$ cDNA), the spliced isoform of *rad8* is stabilized.

Identification and characterization of exon-skipping events

Given the ability of lariat sequencing to detect previously unidentified splicing events, we asked whether exon-skipping events could be detected in *S. pombe*. Because exon-skipping yields a lariat species that *contains* the skipped exon, we looked for peaks that contained an exon along with its flanking introns. More than twenty such peaks were identified (Table 2.6, Methods), including the three shown in Figure 2.3a. As predicted by the lariat sequencing data, PCR using primers that target the exons flanking the skipped exon demonstrate the production of both the canonical and alternatively spliced isoforms of the *srrm1* transcript in wild type (Dbr1⁺) cells. Surprisingly, the exon-skipping event was not readily apparent for either the *alp41* or *qcr10* transcripts in wild type (Dbr1⁺) cells grown under standard conditions. Because our lariat sequencing dataset was generated from a pooled sample of RNAs derived from many environmental stresses (Table 2.1), we asked whether the alternatively spliced isoforms were generated in response to a particular stress. Indeed, the only samples in which the alternative isoform of *alp41* could be detected were those in which the cells had been exposed to heat shock, whereas the alternative isoform of *qcr10* was specifically elicited in response to cold shock (Figures 2.3a, S2.3). A clue as to the mechanism of how cold shock could

induce the skipping of *qcr10* exon 2 is revealed by the predicted secondary structure of the skipped exon unit (intron1-exon2-intron2). The sequence at the 5' end of intron 1 is predicted to form a hairpin with the sequence at the 3' end of intron2 (Fig 2.S9, <http://mfold.rna.albany.edu/>). One model for cold-induced exon skipping would involve formation of this hairpin between introns 1 and 2 in colder temperatures, leading to looping out of exon 2, and failure of that exon to interact with the splicing machinery. Such a model could be tested using mutagenesis to disrupt predicted base pairing, and compensatory mutagenesis to restore base pairing, and assessing cold-induced exon 2 skipping for both strains. Sequencing of the bands corresponding to the alternative products for all three genes confirms the predicted alternative splicing events.

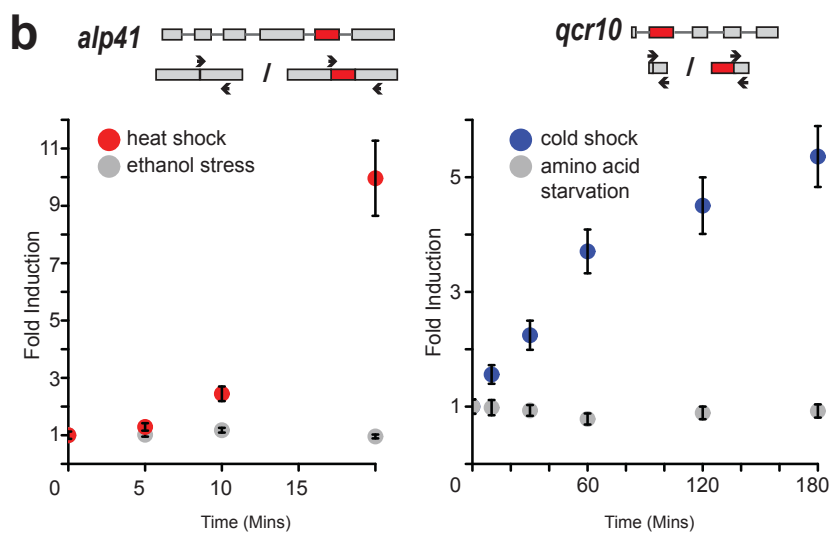
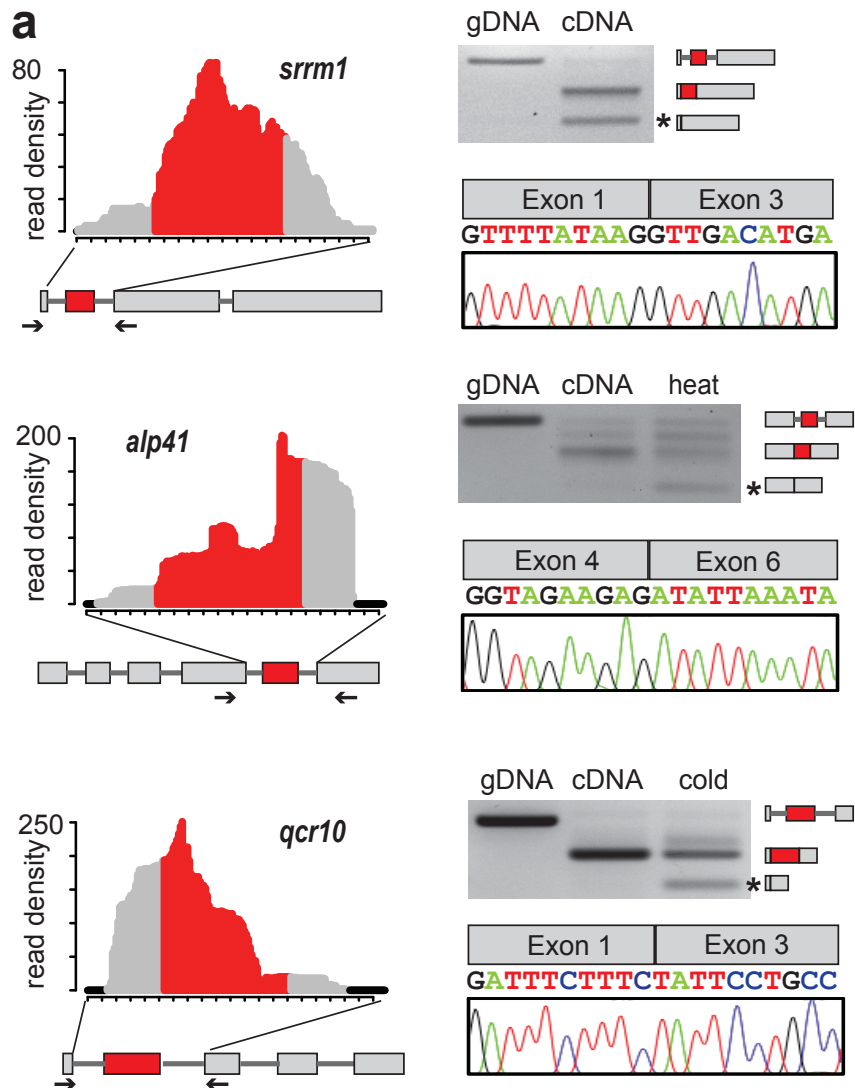


Figure 2.3

Novel instances of constitutive and inducible exon skipping. **(a)** Read density plots surrounding alternatively skipped exons (red) within the *srrm1*, *alp41* and *qcr10* transcripts. Peaks are coloured red where the reads overlap with the skipped exon, and grey where they overlap with the flanking introns. PCR products generated using either genomic DNA or cDNA from wild type cells grown under normal conditions are indicated; cDNA from wild type cells exposed to either heat or cold shock are also indicated. Locations of the unspliced, spliced, and exon-skipped (*) products are noted (see also SI Methods). **(b)** Quantitative PCR using primers that specifically amplify the canonical or alternative isoforms of either *alp41* or *qcr10* demonstrate an increase in exon skipping specifically during the indicated stresses. Values are an average of three biological replicates with three technical replicates each. Error bars represent standard deviation.

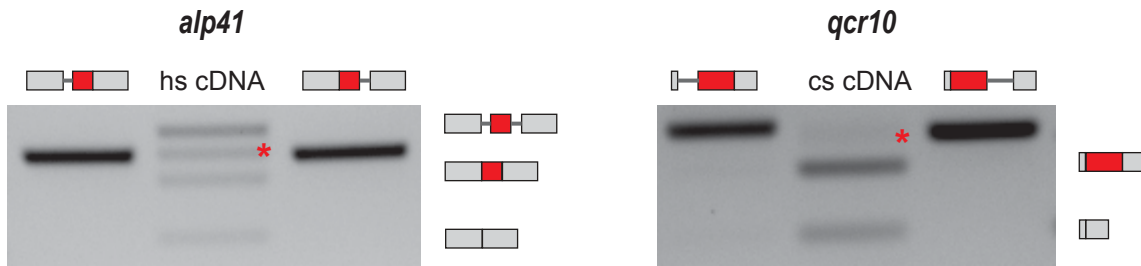


Figure 2.S3

Authentic standards representing versions of the *alp41* and *qcr10* genes where either the upstream or the downstream intron flanking the alternative exon was retained (see Methods). These standards were amplified using the same flanking primers used to amplify cDNA from a heat shock experiment (*alp41*) or from a cold shock experiment (*qcr10*) to reveal the exon skipping events in Figure 3a. These amplified standards were run on a gel flanking the amplified heat shock cDNA for *alp41* and flanking the amplified cold shock cDNA for *qcr10*. Cartoons to the right of each gel depict either the fully unspliced product (*alp41*), the canonically spliced product (*alp41* and *qcr10*) or the exon-skipped product (*alp41* and *qcr10*). Red asterisk depicts the putative intron retention band in the middle experimental sample lanes. Cartoons above each gel represent the upstream retained intron standard (leftmost lane) or downstream retained intron standard (rightmost lane).

| skipped exon | reads | short gene name | Sequencing exp. | RNA-seq reads |
|---------------------|--------------|------------------------|------------------------|----------------------|
| SPBC428.01c exon 2 | 851 | nup107 | short, long | 0 |
| SPAC22F3.05c exon 3 | 663 | alp41 | short, long | 0 |
| SPBC11C11.01 exon 2 | 349 | sr140 | short | 0 |
| SPBP4H10.08 exon 2 | 314 | qcr10 | short, long | 0 |
| SPBC1709.02c exon 2 | 311 | vrs1 | short, long | 0 |
| SPAC22F3.05c exon 5 | 309 | alp41 | short, long | 2 |
| SPCC794.07 exon 2 | 307 | lat1 | long | 0 |
| SPBC1105.09 exon 3 | 158 | ubc15 | short | 0 |
| SPAC27D7.08c exon 6 | 154 | mettl16 | long | 3 |
| SPAC1002.07c exon 3 | 139 | ats1 | short | 0 |
| SPBC1711.03 exon 2 | 120 | aim27 | long | 0 |
| SPAC20G4.04c exon 3 | 107 | hus1 | short, long | 0 |
| SPBC17D11.01 exon 2 | 94 | nep1 | long | 0 |
| SPCC825.05c exon 2 | 89 | srrm1 | long | 2 |
| SPCC1183.05c exon 4 | 85 | lig4 | long | 0 |
| SPBC2D10.15c exon 3 | 83 | pth1 | short, long | 1 |
| SPAC23C4.17 exon 2 | 78 | trm402 | long | 1 |
| SPBC3E7.07c exon 3 | 76 | duf757 | short, long | 0 |
| SPCC126.11c exon 2 | 73 | | short | 5 |
| SPBC4C3.10c exon 2 | 73 | pre3 | short, long | 0 |
| SPCC1840.04 exon 2 | 63 | pca1 | short | 0 |
| SPBC342.04 exon 2 | 63 | rpn1301 | long | 2 |
| SPAC16.01 exon 2 | 61 | rho2 | long | 0 |

Table 2.6 - Exon skipping events deduced from peaks with at least 50 reads

To further characterize the alternative splicing of the *alp41* and *qcr10* transcripts, quantitative PCR primers were designed to specifically detect both the canonical and alternative isoforms of each transcript (Figures 2.S4 and 2.S5, Methods). Production of the alternative isoforms of both transcripts are rapidly induced in response to specific stressors (Figure 2.3b), with the alternative *alp41* isoform showing a ten-fold increase, and the alternative *qcr10* isoform showing a five-fold increase over the course of the experiments. Importantly, no changes in alternative isoform levels are observed for either transcript in response to other stress conditions, indicating that the exon-skipping events reflect specific, regulated responses to distinct stressors.

Given *a priori* knowledge about the specific transcripts that are subject to alternative splicing, we returned to the published *S. pombe* datasets to determine whether any RNA-seq evidence existed for these exon skipping events (29, 30). Importantly, for seven out of the twenty three events we describe here, small numbers of RNA-seq reads can be identified that independently confirm our observations (Figure 2.S6, Table 2.6). The low number of alternative reads and their propensity to be mis-aligned presumably precluded their identification in the original work. While computational approaches for identifying splicing junctions within RNA-seq datasets continue to improve (39, 40), the statistical

limitations inherent to the low read counts associated with these alternative splicing events will complicate their ability to be distinguished from noise.

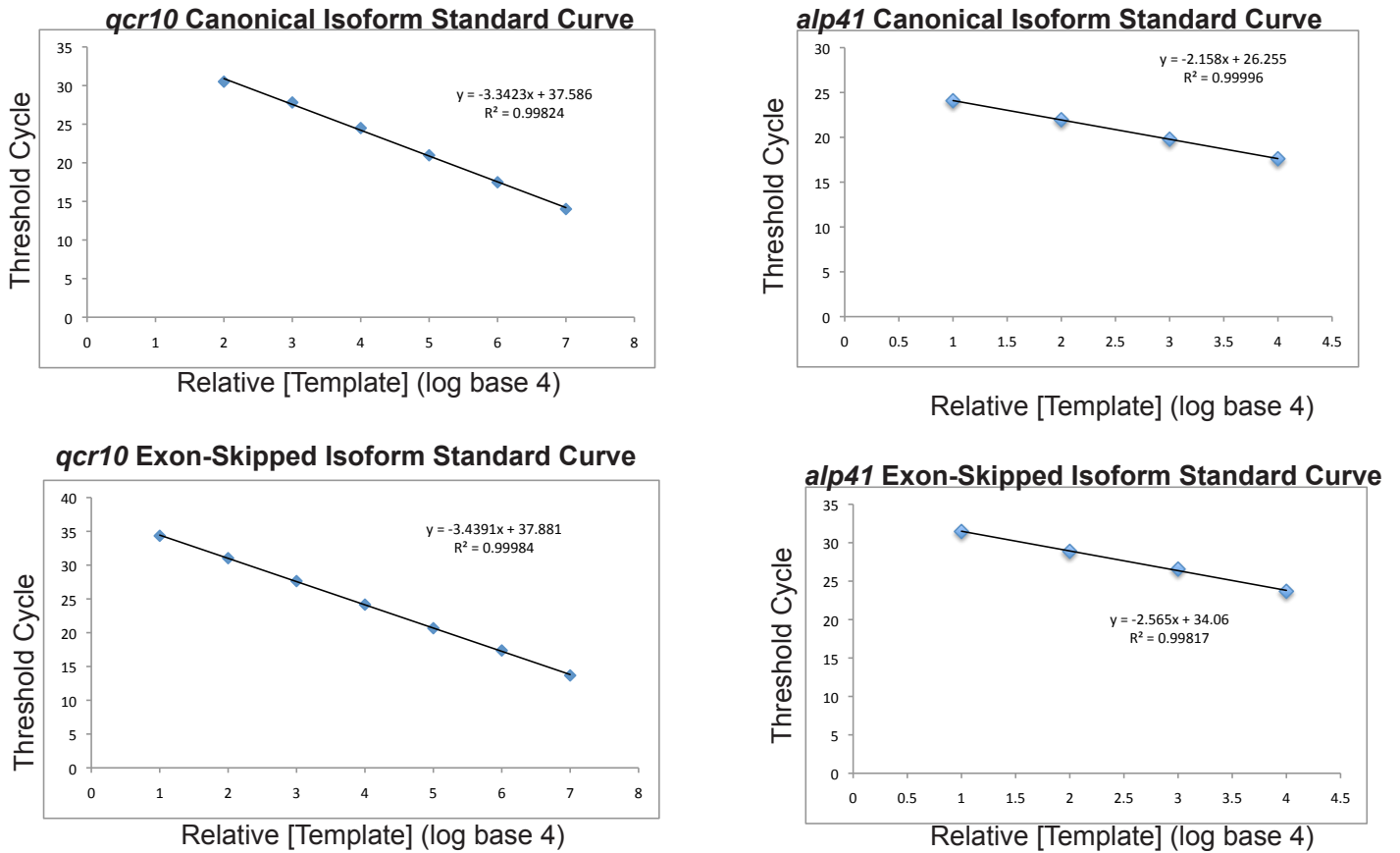
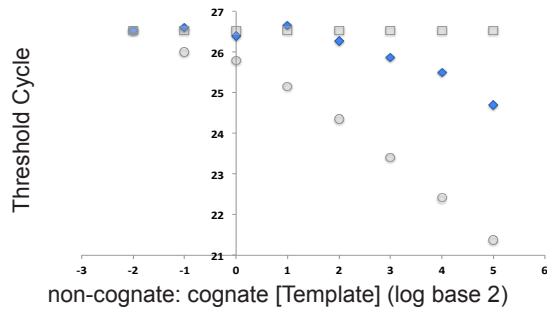


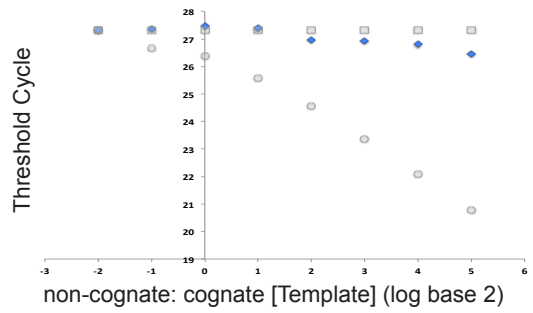
Figure 2.S4

Standard curves generated using qPCR primers targeted to either the canonical or exon-skipped isoform of *alp41* and *qcr10*. The appropriate authentic standard was used as a template in each reaction (see Methods).

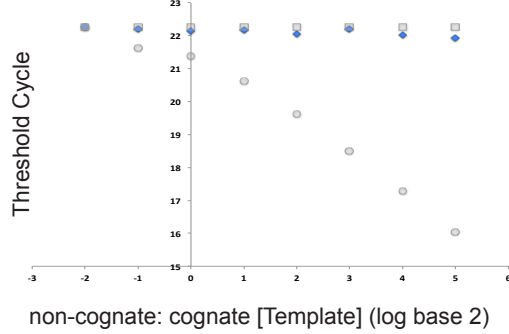
a *qcr10* exon-skipped Isoform primers Specificity Curve



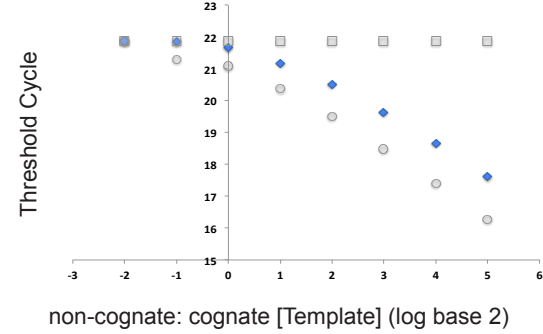
alp41 exon-skipped Isoform primers Specificity Curve



qcr10 Canonical Isoform primers Specificity Curve

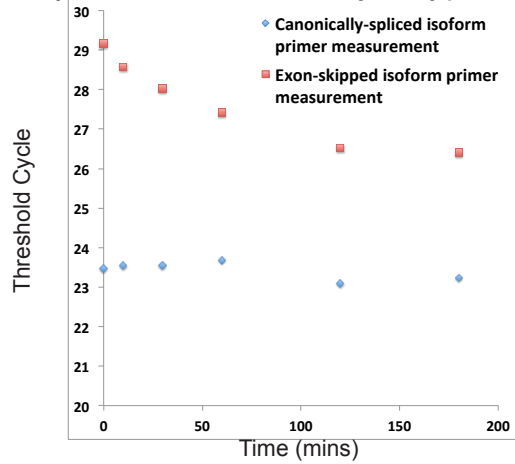


alp41 Canonical Isoform primers Specificity Curve



b

qcr10 Cold Shock threshold cycles by primer pair



alp41 Heat Shock threshold cycles by primer pair

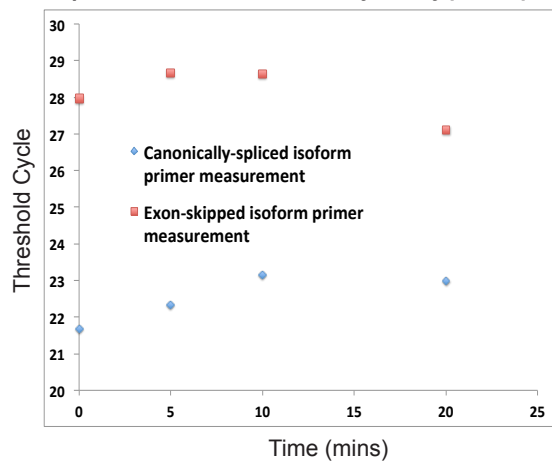
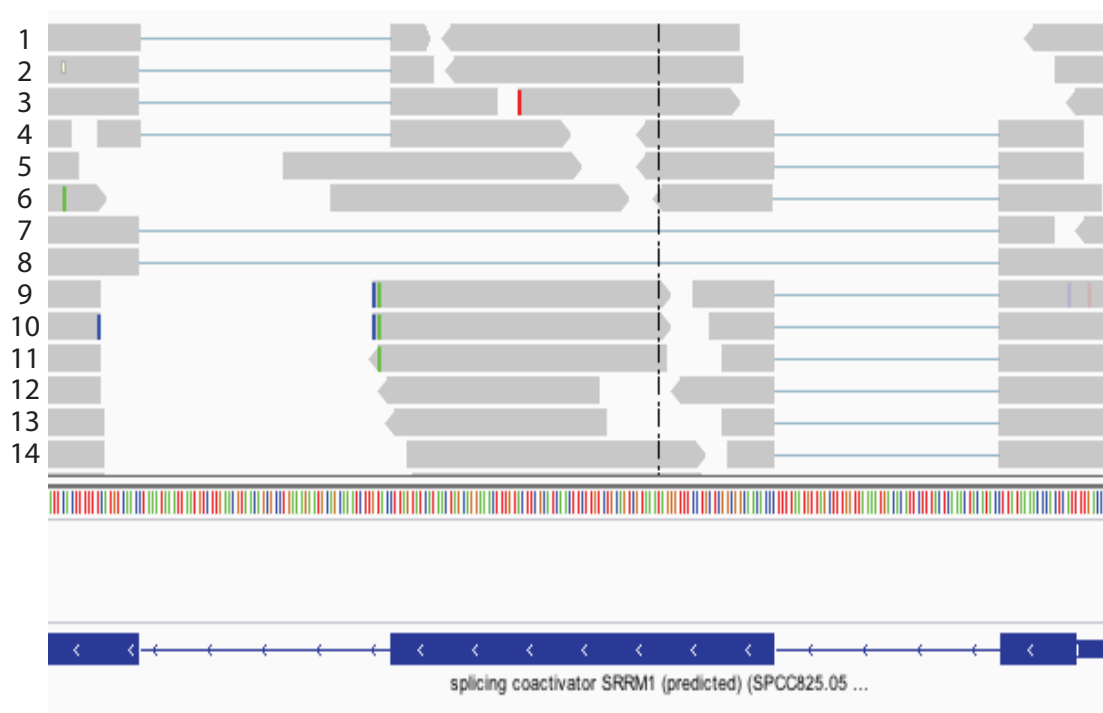


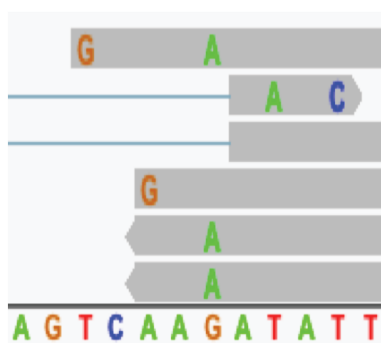
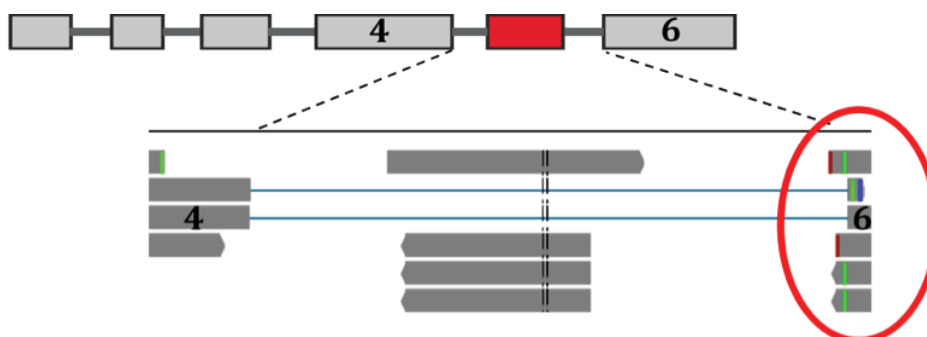
Figure 2.S5

Test of specificity for each of the four primer pairs used in the qPCR experiments shown in Figure 2.3b. **(a)** A test of specificity was carried out using each of the four primer pairs used to detect both the alternative and canonically spliced isoforms of the *qcr10* and *alp41* genes. For each primer pair, qPCR was performed in a reaction with an increasing ratio of non-cognate to cognate template. The amount of cognate template was held constant at a level calculated to give the same threshold cycle as the lowest threshold cycle observed for the primer pair in the experiment in Figure 2.3b. The non-cognate template was doubled over eight steps, starting at a ratio of 1:4 (non-cognate:cognate) and ending at a ratio of 32:1. The results are shown as threshold cycle versus the \log_2 ratio of non-cognate:cognate template (blue diamonds). On each graph are also presented the ‘expected’ results for instances where the primer pairs lack any specificity (can equally detect either the cognate or non-cognate template; open, grey circles), or exhibit perfect specificity (no cross-reactivity with the non-cognate template; open grey boxes). **(b)** The threshold cycles given by both the canonically-spliced isoform primer pair and the exon-skipped isoform primer pair for each time point of the heat shock (*alp41*) and cold shock (*qcr10*) qPCR time course experiments shown in Figure 2.3b.

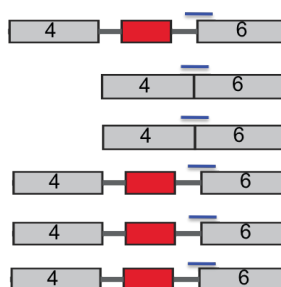
a *srrm1*



b *alp41*



Inferred Alignment



Zero Mismatch Alignment

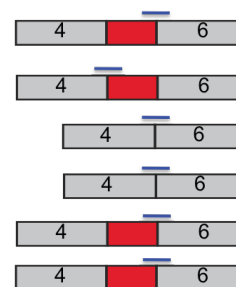
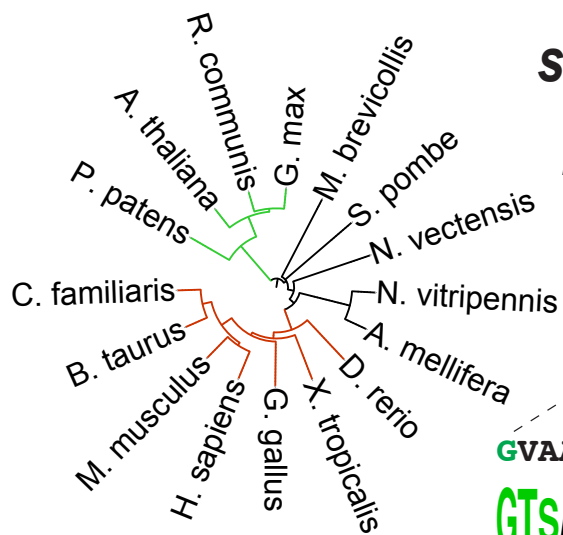


Figure 2.S6.

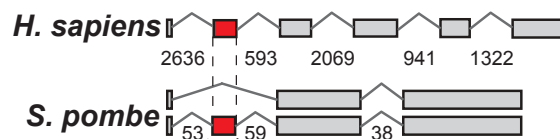
Reanalysis of the BROAD *S. pombe* RNA-seq data reveals a handful of exon-skipping reads, independently confirming the alternative splicing events detected by lariat sequencing. **(a)** A screen-shot from IGV (66) showing RNA-seq reads mapping to *srrm1* in log phase growth, with numbered identifiers added to each line for clarification. Two reads (line numbers 7 and 8) indicate alternative splicing of exon 1 to exon 3, compared with nine reads that indicate the canonical exon1-exon2 splicing event (only four of which are shown on lines 1-4) and fourteen reads that indicate the canonical exon2-exon3 splicing event (only nine of which are shown on lines 4-6, and 9-14). **(b)** Two screen-shots from IGV showing RNA-seq reads mapping to *alp41* in log phase growth, showing either a wide view (upper panel) or a zoomed view (lower panel). The zoomed view is a close-up of the region in the red circle in the wide view. The same subset of reads is presented in the six lines in both panels. In the zoomed view, both the alignment inferred in the original study (left column) and the zero-mismatch, most parsimonious alignment (right column) are shown for each of the six reads. Reads are indicated as blue horizontal lines overlaid on the cartoon of different *alp41* splicing isoforms. For *alp41*, one read correctly indicates splicing of exon 4 to exon 6 (shown on line 3), one read is misaligned as an exon4-exon6 junction read with two mismatches (shown on line 2) whereas it can align perfectly as a canonical exon4-exon5 read; 3 reads are misaligned as intron5–exon6 junction reads with one or two mismatches (shown on lines 1, 5, and 6), whereas they can align perfectly as canonical exon5-exon6 junction reads; and finally one read is mis-aligned as an intron5–exon6 junction read with one mismatch (shown on line 4) whereas it can align perfectly as an alternative exon4-exon6 junction read. For *alp41* in log phase therefore, there are two exon4-exon6 alternative reads, one canonical exon4-exon5 read and three canonical exon 5-exon6 reads.

Evolutionary conservation of alternative splicing events

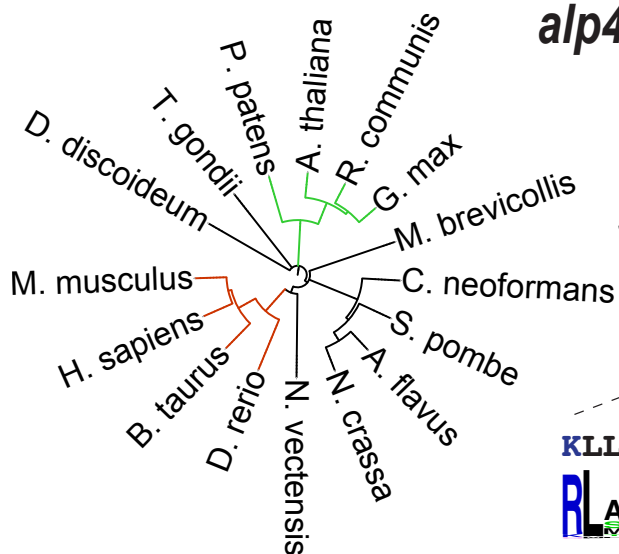
To assess the potential evolutionary conservation of the identified exon-skipping patterns between *S. pombe* and multicellular eukaryotes, we adopted an approach from two previous studies that compared alternative splicing patterns between species from different eukaryotic clades (flowering plants and mosses), where it was argued that a necessary property of a conserved alternative splicing pattern between species is the conservation of the gene architecture that allows such alternative splicing (41, 42). Indeed, it was argued that conservation of gene architecture in and of itself was sufficient to predict the conserved alternative splicing patterns themselves. Accordingly, we examined the global conservation of exon-structure between the *S. pombe* and human genomes. For all 2,641 annotated internal coding exons in the *S. pombe* genome, a search was conducted within the human genome for an orthologous exon, using three criteria to define conservation: the host genes had to be clear orthologs; the exons themselves had to be clear orthologs; and the orthologous exons had to have an identical nucleotide length (Table 2.7*, Methods). Under these metrics, four of the twenty three skipped exons identified by lariat sequencing were conserved, including those in the *srrm1*, *alp41*, *pth1* and *SPAC27D7.08c* genes (Figures 2.4, 2.S7, and 2.S8). Of the remaining 2,618 *S. pombe* internal coding exons, 109 are conserved in humans by these metrics.



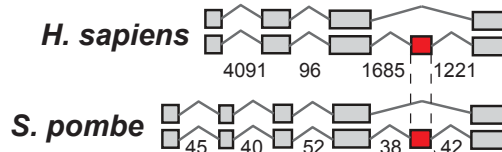
srrm1



GVAAEQETLFTTADKKLMRSTKFPASYDTK
GTSAEQDNRFSENKQKLLKQKFAECLKK



alp41



KLLFTSILVLANKSDVSGALSSEEISK
RLAGASLLYFANKQDLGALSTEEAEK

Figure 2.4

Evolutionary conservation of skipped exons in the *srrm1* and *alp41* genes of *S. pombe*. Left: pruned phylogenetic trees depicting model species that have an ortholog of the skipped *S. pombe* exons in *srrm1* (top) or *alp41* (bottom), drawn according to NCBI taxonomy relationships (60). Full trees are shown in Figure 2.S8. Vertebrate branches are coloured red and plant branches are coloured green. Right: For the *srrm1* and *alp41* genes, the two alternate *S. pombe* isoforms corresponding to the skipping events are shown with the corresponding isoforms from Ensembl (human) or UniProt (mouse) shown above. Only the first six exons of the human ortholog of the *srrm1* gene are shown. The evolutionarily conserved exons are shown in red, and intron lengths are indicated in the longer isoform. Below the gene diagrams are shown the peptide translations of the *S. pombe* skipped exons and below that a peptide motif constructed from a multiple sequence alignment of the translated orthologous exons from all species for which an orthologous exon was found (59).

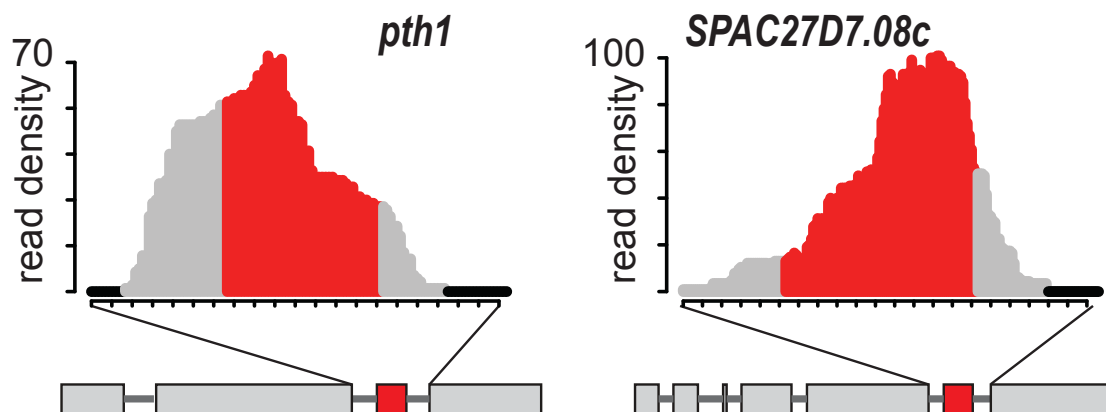
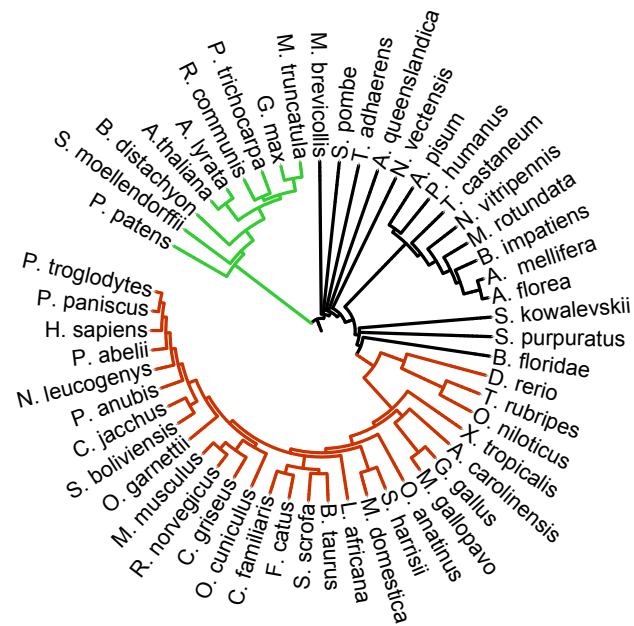
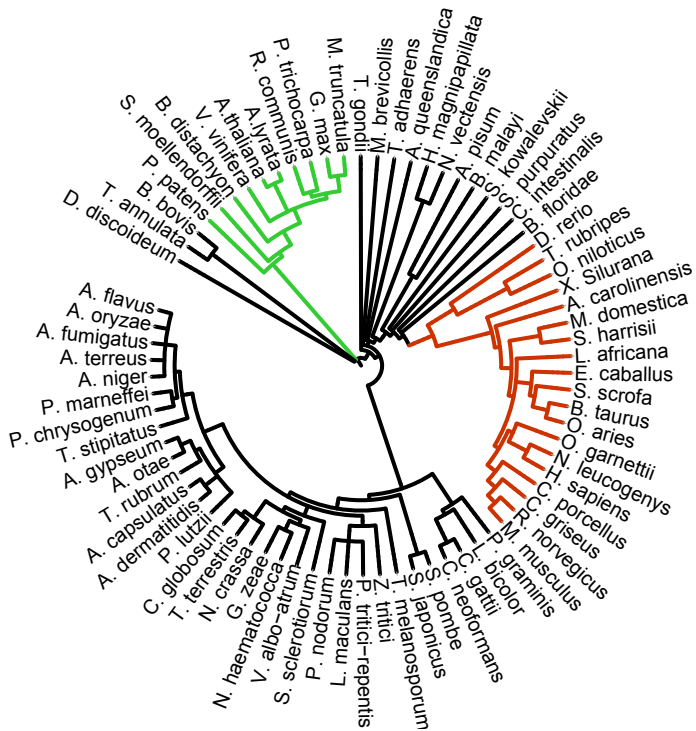
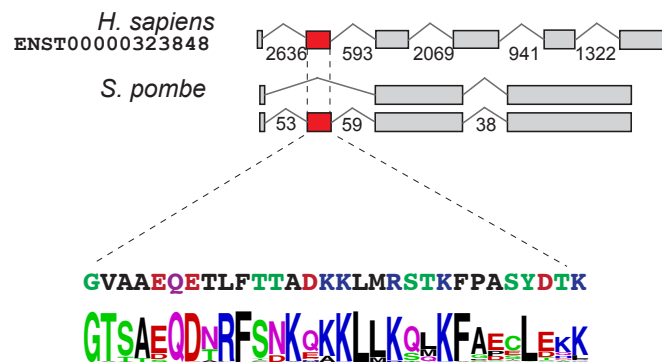


Figure 2.S7

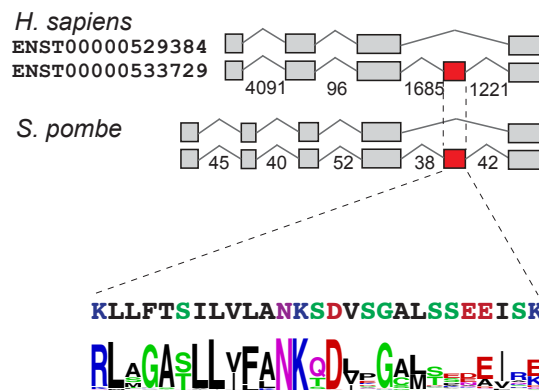
Lariat sequencing read densities for exon skipping events in the *pth1* and *SPAC27D7.08c* genes.

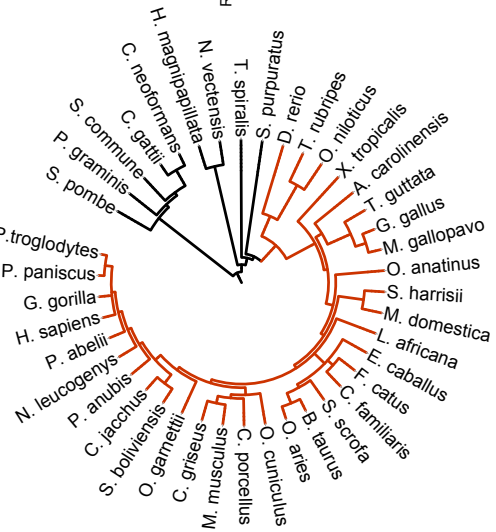
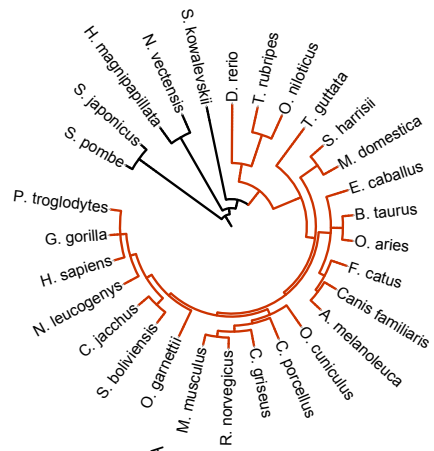


srrm1



alp41





M. musculus

Q9CQG2-2

Q9CQG2-1

10kb 3297 4507 3663 7706 1352 12kb 115

S. pombe

46 78 46 48 47 56

ISNYGIYELALGKTKRWIICWSFQAMRPHN

VPKVT_YTEFCQGRTRWALAWSFYDDV_TVP

Figure 2.S8

Evolutionary conservation of the skipped exons in the *srrm1*, *alp41*, *pth1*, and *SPAC27D7.08c* genes. Left: full phylogenetic trees depicting all species that have an ortholog of the skipped *S. pombe* exons in *srrm1*, *alp41*, *pth1* or *SPAC27D7.08c* (top to bottom, see SI Methods), drawn according to NCBI taxonomy relationships (8). Vertebrate branches are coloured red and plant branches are coloured green. Right: For the *srrm1*, *alp41*, *pth1* or *SPAC27D7.08c* (top to bottom) genes, the two alternate *S. pombe* isoforms corresponding to the skipping event are shown with the corresponding isoforms from Ensembl (human) or UniProt (mouse) shown above. Only the first six exons of the human ortholog of the *srrm1* gene are shown. The evolutionarily conserved exons are shown in red, and intron lengths are indicated in the longer isoform. Below the gene diagrams is shown the peptide translation of the *S. pombe* skipped exon and below that a peptide motif constructed from a multiple sequence alignment of the translated orthologous exon from all species for which an orthologous exon was found (9).

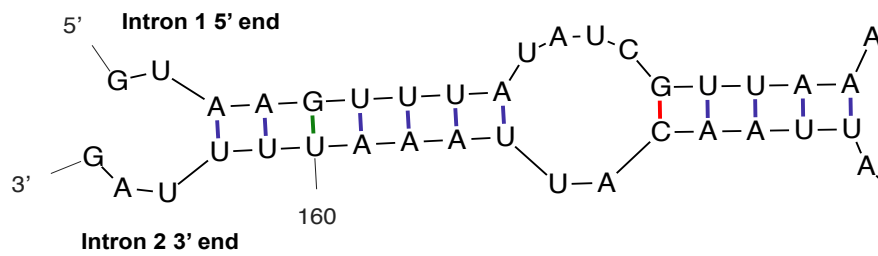


Figure 2.S9

Predicted hairpin formed between the 5' end of *qcr10* intron 1 and the 3' end of intron 2, looping out exon 2.

While conserved gene architecture is a necessary condition for a conserved alternative splicing pattern between two species and has previously been argued as indirect evidence of conserved alternative splicing, the gold standard is undoubtedly direct experimental evidence of alternative splicing in both species. Remarkably, the skipped *S. pombe* exons of three of the four genes noted above are not only structurally conserved with their human orthologs, but alternative skipping of the orthologous exons have been observed within their mammalian counterparts (Figures 2.4 and 2.S8). For the *alp41* and *pth1* genes the human orthologs have been demonstrated to skip their orthologous exons, in each case generating protein-coding alternative isoforms (43, 44). Moreover, a recent RNA-seq experiment in macaque provides further evidence for the skipping of these exons (ref. 45; Methods). Similarly, for the *SPAC27D7.08c* gene, the orthologous mouse exon is subject to exon skipping (45, 46). Interestingly, an analysis of RNA-seq data from the Illumina BodyMap study reveals tissue-specific regulation of the alternative isoform of the human ortholog of *alp41* (GEO accession GSE30611, Methods), further augmenting the relationship with the regulated behaviour of the *S. pombe* ortholog demonstrated here.

To better understand the evolutionary history of these conserved exon skipping events, a systematic search of all sequenced eukaryotic genomes from the NCBI was carried out (Methods). This endeavour yielded over one hundred

species in which one or more of these exons were stringently conserved (Figures 2.4 and 2.S8, Table 2.8*). Notably, the skipped exon of *alp41* is conserved by the previously-defined metrics in vertebrates, plants, and fungi, and in all but two cases the length of the exon is identical. This evolutionary analysis suggests that the gene structure necessary to support alternative splicing of this exon is at least as old as the common ancestor of animals, plants and fungi. Although no evidence currently exists for skipping of the conserved exon in these species, transcriptome data in these species are markedly lower than that available for humans.

While the data presented above support an ancient origin for the alternative splicing in *alp41* and *pth1*, we sought to test the alternative hypothesis that the identical alternative patterns seen for these genes could be the result of convergent evolution. To do so, we asked, if a random sample of 23 internal coding exons was selected, how often we would expect to observe at least two *S. pombe* exons that have orthologous exons in humans, and for which the human orthologous exon is skipped. The results indicate that the two observed exons are significantly more than expected by chance under the convergent evolution model (Methods), and that this significance is robust to relaxed definitions of exon conservation compared to those used to identify the orthologous exons in the *alp41* and *pth1* genes. Such a finding is consistent with the idea that *S. pombe* is undergoing loss of exon skipping concomitant with

general intron loss (47), and that those exon skipping events that remain are relatively ancient, and not recent innovations.

Unlike the previous examples, there is no evidence for exon skipping of the conserved exon in the human ortholog of *srrm1*. However, *srrm1* gene structure is specifically conserved between vertebrates, insects, plants and fungi exclusively at this exon and the two flanking exons (Figures 2.4 and 2.S8). Importantly, the *srrm1* protein is an SR-like protein that functions to regulate alternative splicing in higher eukaryotes (48). The skipped exon of *srrm1* encodes a stretch of basic amino acids just upstream of the PWI-domain demonstrated to play an important role in RNA binding (49), suggesting that the alternative protein isoform has dramatically different RNA-binding properties. Notably, this separation of basic motif (in exon 2) from PWI motif (at the start of exon 3) is conserved in vertebrate *srrm1* genes, while the rest of the gene is more variable. There are well-documented examples in higher eukaryotes where transcripts encoding SR-proteins are themselves subject to regulation via alternative splicing (50, 51), suggesting the intriguing possibility that alternative splicing of the *S. pombe srrm1* mRNA may function to regulate the levels of functional Srrm1 protein.

2.3 Discussion

Here we describe the development and implementation of a technique that leverages sequencing of the lariat introns that are excised during splicing to enable high-sensitivity detection of genome-wide pre-mRNA splicing events. This lariat sequencing approach was applied to the unicellular yeast, *Schizosaccharomyces pombe*, an organism that shares with vertebrates many of the features of alternative splicing, yet for which recent RNA-seq experiments had failed to detect instances of exon-skipping (29, 30). Lariat sequencing was able to identify hundreds of previously uncharacterized splicing events in *S. pombe*, including over twenty examples of exon-skipping, several of which are shown in Figure 3. While exon skipping has been previously demonstrated at similarly low frequency in other unicellular organisms (15–17), these events represent the first examples of this type of alternative splicing in this organism. More importantly, however, is the demonstration here that several of the *S. pombe* exons that are subject to skipping are highly conserved across eukaryotic species spanning several major eukaryotic lineages, and that experimental evidence exists for the alternative splicing of the orthologous exons in mammalian species. To our knowledge, these exon skipping events represent the first known examples of specific alternative splicing patterns that are identical between species from different eukaryotic kingdoms for which both

isoforms encode functional proteins, and for which conditional regulation has been demonstrated. These findings provide strong support for the hypothesis that functional alternative splicing predates multicellularity (13).

Importantly, exon-skipping in multicellular eukaryotes is thought to be facilitated via an ‘exon-definition’ model of spliceosome assembly, yet mechanistic studies of splicing in *S. pombe* suggest that assembly occurs through the ‘intron-definition’ model more commonly associated with unicellular eukaryotes (52). Nevertheless, it remains unclear whether spliceosome assembly on select introns in *S. pombe* can occur by exon-definition; indeed, it has been proposed that both intron- and exon-definition exist in *D. melanogaster* (53), and assembly via intron- and exon-definition can occur on a single human transcript (54). An important future goal will be to determine whether the mechanistic bases by which these alternative splicing events are regulated are indeed conserved between *S. pombe* and humans. The highly tractable genetics of *S. pombe* provide a powerful system in which to undertake these experiments.

Beyond the identification of alternative splicing events, lariat sequencing was also able to identify over two hundred previously uncharacterized splicing events within the *S. pombe* genome, many of which appear to be canonical, constitutively spliced introns. In fact, because the current experiments were designed primarily to recover the larger lariats that result from exon-skipping events, the novel splicing events found in this work almost certainly represent

only a portion the unannotated splicing events that occur in *S. pombe*. Thousands of currently annotated introns in the *S. pombe* genome have lengths that are shorter than 100 nucleotides; a simple extrapolation based on the ratio of novel versus annotated introns detected in this experiment suggests that as many as one thousand splicing events are likely to exist within the *S. pombe* genome that have yet to be identified.

The majority of the novel splicing events in coding exons of protein coding genes identified in this work result in the production of a transcript that is predicted to be targeted for degradation. It is generally accepted that many cellular mRNAs are subject to rapid cellular destabilization as a mechanism for regulating their gene expression (4). The ability to detect these transcripts, or splicing events within them, in an RNA-seq experiment will be compromised relative to more stable mRNAs. While little is known about the degradation rates of specific lariats (55), lariat intron decay is likely subject to less transcript specific regulation than is mRNA decay, making lariat sequencing a particularly sensitive method for detecting splicing events for which the mature transcript is subject to rapid degradation.

The results presented here have important implications not only for our understanding of splicing in *S. pombe*, but more generally for the use of lariat sequencing as a complement to RNA-seq for characterizing genome-wide splicing events in any organism. The relatively modest sized lariat sequencing

experiment presented here surveyed the combined RNA from dozens of different stressed samples yet was able to detect hundreds of splicing events that went undetected in an RNA-seq experiment with nearly 20 times the sequencing depth. The reasons for this enrichment are almost certainly both technical as well as biological in nature. In fact, for some of the novel introns presented here, limited numbers of RNA-seq reads can be identified that confirm these splicing events, highlighting the bioinformatic challenge associated with *de novo* identification of splicing events (56). Others have characterized the statistical challenge involved in detecting short transcripts in an RNA-seq experiment (57): in essence, an exon-junction is detected in RNA-seq when a read aligns precisely to that junction, which can only happen at a limited number of positions that is directly proportional to the read length. Thus, in RNA-seq, exon-exon junctions behave as if they were individual transcripts with a size approximating the sequencing read length. To be sure, this short-coming of RNA-seq will diminish in scope as newer technologies allow for increased read lengths. Likewise, reductions in sequencing costs will enable RNA-seq datasets to achieve greater depths. However, even with improved read coverage, splicing events that produce unstable transcripts will be better sampled by approaches such as lariat sequencing which capture and sequence the intronic material excised by the splicing machinery. Indeed, the recent identification of one lariat RNA read for every one million reads of

pooled human RNA-seq data demonstrates the potential of this approach (58). Further optimizations of the stabilization, purification, and sequencing of lariats should provide a powerful addition to the arsenal of tools for characterizing splicing within genomes and transcriptomes.

2.4 Materials and Methods

Strains Used

All experiments were performed using one of four strains.

1. **wild type 972 h⁻** from ATCC, # 38366
(<http://www.atcc.org/ATCCAdvancedCatalogSearch/ProductDetails/tabid/452/Default.aspx?ATCCNum=38366&Template=fungiYeast>)
2. ***Δdbr1*** DBR1::Nat, constructed using standard cloning techniques from the wild type strain above.
3. ***Δsmd3*** from Bioneer Haploid Deletion Mutant Set (<http://pombe.bioneer.com>)
4. ***Δupf1*** from Bioneer Haploid Deletion Mutant Set (<http://pombe.bioneer.com>)

Growth Conditions of *Δdbr1* Cells

In order to elicit a broad spectrum of transcriptional and splicing paradigms, the *Δdbr1* strain was exposed to a variety of different growth conditions, with samples collected at a variety of times after exposure to environmental stressors. The growth conditions (summarized in Table 2.1) for *Δdbr1* cells used to make the lariat sequencing RNA pool were as follows.

- A) YES vegetative growth (samples #1-#7): rich liquid media (YES) was made according to the recipe in Forsburg and Rhind (61). For samples 1-4, cells were grown in YES at 30°C overnight to saturation in a 5ml culture, then back-diluted to an OD of 0.05. When cells reached ODs of 0.1, 0.6, 1 and 5, 10 ml of cell culture was harvested by filtration, then placed in a 15mL falcon tube and flash frozen using liquid nitrogen and stored until RNA isolation (see below). For samples #5-7, the

- process was the same, except that cells were grown at the temperatures indicated in Table 2.1, and harvested at an OD of 0.6.
- B) Heat shock (samples #8, #9): cells were grown overnight in YES at a temperature of 25°C then back-diluted to an OD of 0.05. When the cells reached an OD of 0.6 they were shifted to a temperature of 42°C. At 10 and 60 minutes after the shift, cells were harvested as described above.
- C) Cold shock (samples #10, #11): cells were grown overnight in YES at a temperature of 30°C then back-diluted to an OD of 0.05. When the cells reached an OD of 0.6 they were shifted to a temperature of 16°C. At 30 and 180 minutes after the shift, cells were harvested as described above.
- D) Salt, DNA damage, oxidation and other stresses (samples #12-#25): Cells were grown overnight in YES at a temperature of 30°C, then back-diluted to an OD of 0.05. For each stressor (KCl, 4NQO, MMS, Lithium, Ethanol, DTT, H₂O₂), when cells reached an OD of 0.6, pre-warmed YES containing the stressor was added to the cell culture, to give the final concentrations of stressor indicated in Table 2.1. After 20 and 120 minutes of stress, cells were harvested as described above.
- E) Growth in minimal media, EMM, (samples #26-#32): EMM was made according to standard techniques (61). Samples were harvested as described for samples #1-#7 above, except that cells were grown in EMM instead of YES.
- F) Amino acid starvation (samples #33, #34): Cells were grown overnight in EMM at a temperature of 30°C, then back-diluted to an OD of 0.05. When cells reached an OD of 0.6, pre-warmed EMM containing 0.5M 3-amino triazole (3-AT) was added to give a final concentration of 0.05M. After 30 and 180 minutes of exposure to 3-AT, cells were harvested as described above.

- G) Glucose, Nitrogen and Phosphorous starvation (samples #36-#40): For each of these three stresses, cells were grown in EMM at 30°C, then back-diluted to an OD of 0.05. When cells reached an OD of 0.6, 50mL of culture was filter-collected, then resuspended in 50ml of pre-warmed EMM lacking either glucose, nitrogen or phosphorous. This media was prepared according to standard techniques, but omitting either glucose, NH₄Cl or Na₂HPO₄, respectively. After this, cells were harvested at times of 20 minutes and 120 minutes, as described above.
- H) Solid YES agar (sample #41): Cells were streaked to single colonies from a glycerol stock onto a plate of YES agar (YES + 2% w/v Difco Bacto Agar), and grown for 2 days at 30°C. A single colony was transferred to a 15mL falcon tube and flash frozen using liquid nitrogen.

For every other experiment in the paper (all of the flanking PCR experiments in Figures 2.1, 2.2 and 2.3 with wild type *S. pombe*) except the qPCR experiments for the *alp41* and *qcr10* gene (see below), unless otherwise noted, cells of the relevant strain were grown overnight in YES then back-diluted to an OD of 0.05., allowed to grow to OD 0.6, and then harvested by filtration.

Total RNA Isolation from *Adbr1* Cells

From the cell pellets collected above, total cellular RNA was isolated using hot phenol chloroform extraction as follows. Into the 15mL falcon tube containing filter-collected cells was added 2mL Acid Phenol (pH<5.5), followed by 2mL AES Buffer (50mM sodium acetate (pH 5.3), 10mM EDTA, 1% SDS) and the tube was vortexed for 10 seconds, then incubated at 65°C for 7 minutes, vortexing for at least 3 seconds at one minute intervals. The tube was then incubated on ice for five minutes, then the entire mixture was transferred

to a 15mL PhaseLock Heavy Gel tube (5PRIME) and centrifuged at 4°C at 5250 \times g for 5 minutes. 2mL phenol:chloroform:iaa (25:24:1) was then added to the supernatant, and mixed by shaking the tube up and down vigorously for 5 seconds. The tube was then centrifuged again in the same manner, and 2mL Chloroform was added to the supernatant, mixing as before. The tube was centrifuged as before, and this time the supernatant was transferred to a new 15mL Falcon tube, to which 2mL isopropanol and 200 μ L 3M sodium acetate (pH5.3) was added. The resulting solution was mixed by vortexing, and 2mL was transferred to a 2mL centrifuge tube. The 2mL tube was centrifuged for 20 minutes at 4°C at 18,000 \times g, then supernatant was removed by decanting, and 2mL 70% Ethanol was used to wash the RNA pellet. The tube was then centrifuged for 5 minutes at 4°C at 18,000 \times g and the 70% Ethanol wash step and 5 minute centrifugation step was repeated. This second 2mL of 70% Ethanol was then decanted, and the sample was dried using vacuum centrifugation at room temperature. The RNA pellet was resuspended in RNase free water. After isolation, 21 μ g of RNA from each of the 41 samples was pooled together. This sample was split into two 400 μ g aliquots for purification on two-dimensional PAGE – one for the “short” gel conditions and one for the “long” gel conditions

Lariat RNA isolation via Two-Dimensional PAGE

Total cellular RNA from *Adbr1* strains grown under a variety of conditions was isolated as described in Methods. Two different sets of gel running conditions were used in order to optimize the recovery of lariat RNAs

of different lengths from these samples. To recover lariats that were as short as ~100 nucleotides the RNA was initially run in a single lane on a 7.5%, 8.3M urea gel in 1X TBE at 40 mA until the bromophenol blue had migrated 12 cm. This lane was cut out of the gel and recast across the top of a 15% gel. This second gel was run at 30 mA until the bromophenol blue had migrated 28 cm. The lariat RNAs recovered from this 2D gel sample are referred to as “short” lariats in Table 2.6. To ensure recovery of longer lariats, a separate purification was performed where the first dimension utilized a 7.5% acrylamide gel run at 40mA until the xylene cyanol had migrated 31cm, while the second dimension utilized a 20% gel run at 28mA until the xylene cyanol had migrated 68cm. The lariats recovered from this gel are referred to as “long” lariats in Table 2.6. The downstream processing of the RNAs isolated from both sets of gel conditions was identical, as described in Methods. Of the twenty three exon skipping events identified, five were identified exclusively from the ‘short’ gel running conditions, nine exclusively from the ‘long’ gel running conditions, and nine from both.

Library preparation and sequencing.

Library prep was similar to that outlined in Quail *et al.* (62), with modifications as described below.

cDNA synthesis

Gel-purified lariat RNAs were converted into cDNA without prior debranching by using random nonamers as primers in a 20 μ L reaction containing 5 μ g of primer, 1 μ g RNA sample, 50mM Tris-HCl (pH8.5), 3mM MgCl₂, 10mM DTT, 0.2mM each dNTP, and 60ng of MMLV reverse transcriptase. Reactions were incubated overnight at 42°C.

Second strand synthesis

Second strand cDNA was generated using *E. coli* DNA Pol I and RNaseH (Invitrogen) in a 100 μ L reaction containing 20 μ L of the cDNA synthesis mixture, 12 μ M each dNTP, 1X DNA Pol I buffer, 50 Units of DNA Pol I, and 1 Unit of RNase H. The reaction was incubated at 16°C for one hour.

Cleanup

The products of the second strand synthesis reactions were phenol-chloroform extracted, chloroform extracted twice, ethanol precipitated, washed and finally desalted using BioRad Micro Bio-Spin 6 columns equilibrated with water. This extraction, precipitation and desalting clean-up regimen was used after each subsequent step in library preparation.

End Repair

A 100 μ L reaction was set up using 34 μ L dsDNA (volume of second strand reaction left after cleanup), 10 μ L NEB end repair buffer, 5 μ L NEB end repair

enzyme mix and 51µL water. The reaction was incubated at 20°C for 30 minutes. Cleanup was as above (eluted into 34µL).

A-tailing

A 50µL reaction was set up using 32µL DNA (from post-cleanup of end repair reaction), 5µL of 10X NEB Buffer 2, 10µL of 10mM dATP, and 5000 units Klenow exo- (NEB). The reaction was incubated at 37°C for 30 minutes. Cleanup was as described above, but eluted into 22µL water.

Adapter ligation

A 50µL adapter ligation reaction was set up using 19.5µL DNA (from post-cleanup of a-tailing reaction), 25µL DNA ligase buffer, 0.5µL sequencing adapter and 3000units of T4 DNA ligase (Enzymatics). The reaction was incubated at room temperature for 15 minutes.

PCR Amplification

Adapter-ligated DNA was purified on a 6% acrylamide gel. The gel was cut between 150 and 300 nucleotides, and DNA isolated from the gel was precipitated and subjected to Phusion PCR as follows. Reaction mix was 100uL, with 4uL DNA sample precipitated from gel, 1X HF buffer, dNTPs (0.3mM each), 0.25µM long Solexa PCR primer, 0.25µM short Solexa PCR primer, and Phusion polymerase (NEB). The PCR protocol was 98°C for 30 seconds, then 24 cycles of 98°C for 10 seconds, 65°C for 30 seconds, 72°C for

30 seconds. Final extension was 72°C for 5 minutes. The PCR product was purified on a 6% acrylamide gel, recovering the material above the primer dimer (100 bp). The purified products were sequenced on an Illumina GAIIx machine platform.

Intron Discovery

Sequence reads were initially aligned to the *S. pombe* reference genome from the Sanger centre website using bowtie version 0.10.1, allowing for zero mismatches, and alignments at up to three different genomic regions. The decision to allow reads that could align to multiple genomic loci read was made to allow alignment to ribosomal RNAs, which are often present in multiple genomic copies. To determine the number and percentage of reads aligning to different types of genomic features, aligned read positions were compared to the pombe_040211.gff file from the Sanger centre website. A second, more conservative, alignment was performed that only permitted unique read alignments; this alignment was used for all subsequent analyses.

After read alignment, “peaks” were defined as non-zero read regions with at least five reads. Peak edges were defined as the nucleotide on either side of the peak where the read count became zero. A total of 1541 peaks overlapped a single intron annotated in the *S. pombe* gff files, of which 1325 had at least 90%

of peak density contained within the intron (Table 2.4). For nine of the 1541 detected introns, a small number of reads were detected that extended beyond the annotated boundaries. For example, the peak for *rpl4301* shown in Figure 2.1 contains 237,155 reads that map entirely within the intron, but an additional twelve reads that map to the exons flanking the intron. We presume that these reflect a small amount of contamination of linear RNA since these introns are all contained within very highly expressed genes, but cannot rule out the possibilities that they reflect either: (1) aberrant splicing at non-canonical splice sites within this transcript; or (2) the isolation of lariat-intermediates of the splicing reaction. In Figure 2.1, only the 237,155 reads that mapped to the intron are shown, with the remaining 12 reads having been excluded. Of the nine peaks showing this behaviour, all had at least 10,000 reads that mapped within the intron, and none had more than 12 reads mapping outside of the intron. None of the other peaks displayed in any of the figures were trimmed.

Whereas a minimum of five reads were required for peak calling, only peaks with at least fifty reads were further considered as candidate introns. For each of these peaks, thirty base sequence windows both up- and downstream of each of the peak boundary coordinates were scanned for potential splice site sequences. The upstream window was scanned for all GT dinucleotides, representing possible 5' splice sites. The downstream window was scanned for

all possible branch point sequences and all AG dinucleotides, representing possible 3' splice sites. For every peak, these sequences were scanned separately in both the Watson and the Crick strand. For each peak, four different intronic features were then considered: 5' splice site, 3' splice site, branch point sequence, and branch point to 3' splice site distance. Each of these was scored by calculating the log-odds ratio under a foreground versus background probability model. For the 5' splice site, the foreground model was a feature-specific first-order Markov model (as detailed in Lim and Burge (63)), built using the first 8 nucleotides of every intron annotated in the pombe_040211.gff file (see above). A background model in the form of a dinucleotide frequency matrix was built using exonic sequences and intronic sequences that were not part of the first 8 nucleotides of the intron. The score of any putative 5' splice site sequence was calculated as the log odds ratio of that 8-mer under the 5' splice site model relative to the background model. Putative 3' splice site sequences were scored analogously to the 5' splice sites, considering instead the last five nucleotides of the intron. Branch point sequences were scored using branch point sequences from the "aligned_introns" file from the Sanger centre to build the foreground model, and all exonic and intronic sequence specified by the pombe_040211.gff file to build the background model. Finally, the branch point to 3' splice site distance was scored, using for the foreground model an empirical probability distribution (smoothed by a Gaussian kernel) of

the number of nucleotides from the branch point adenosine to the terminal nucleotide of the 3' splice site, extracted from the “aligned_introns” file, and for the background model a uniform probability distribution, with minimum and maximum values corresponding to the minimum and maximum empirical values for branch point adenosine to terminal 3' splice site nucleotide observed in the “aligned_introns” file. To get a final score for each putative intron, an unweighted sum was taken of the four scores described above. The scoring process was carried out in both Watson and Crick strands, and the higher final score was chosen as the putative splicing event.

Classifying Novel Introns

Novel splicing events were classified based on overlap with known genomic features, specified by the pombe_09052011.gff file from the Sanger Centre website, and the SP2_CALLGENES_FINAL_3.gff3 and SP2_ncRNA.gff3 files from the BROAD Institute website. Intron type categories were then annotated as follows:

Exon Skipping: Peak completely spans an exon, starting in the upstream intron and ending in the downstream intron

5'UTR: Intron is completely contained within or partially overlaps with the annotated 5'UTR of a gene, on the same strand.

3'UTR: Intron is completely contained within or partially overlaps with the annotated 3'UTR of a gene, on the same strand.

cds in-frame: Intron is completely contained within a coding exon of an annotated gene, on the same strand, and the intron has a length that is a multiple of three.

cds frameshift: Intron is completely contained within a coding exon of an annotated gene, on the same strand, and the intron has a length that is not a multiple of three.

5' modelchange: Intron spans the annotated start codon of a gene, on the same strand.

3' modelchange: Intron spans the annotated stop codon of a gene, on the same strand.

Antisense: Intron completely overlaps coding or untranslated region of a protein-coding gene, but on the opposite strand.

Intronic snoRNA: Intron contains an annotated snoRNA, on the same strand

snoRNA intron: Intron is contained within an annotated snoRNA, on the same strand.

ncRNA: Intron is contained within a non-coding RNA on the same strand, and does not overlap a protein-coding gene or snoRNA on either strand.

unclassified: Intron does not overlap any features in either strand.

Determining the percentage of single exon-skipping events covered by lariat sequencing

The length of all potential single exon skipping events was determined as follows. For each non-terminal exon in the *S. pombe* genome, the length of the exon was added to the sum of the lengths of the two flanking introns. A total of 2808 intron-exon-intron units (potential single exon-skipping events) were identified with the shortest being 40 nucleotides and the longest being 4015 nucleotides in length. Out of this total, 1853 such units (66%) had lengths in the range of 85 – 400 nucleotides, the size range captured by our gel conditions.

Characterizing the slow-migrating bands in Figure 2.3a

In the gels for *alp41* and *qcr10* in Figure 2.3a, bands are apparent in the cDNA template lanes that run between the fully unspliced products and the constitutively spliced products (see also Figure 2.S3). We tested the idea that these bands represented retention of either the up- or down-stream intron that flanks the internal exon, due to either intron retention or incomplete splicing.

To this end, authentic DNA amplicons were constructed that recapitulated either the retained upstream or downstream intron, using chimeric primers to construct the intron deletions. For example, to make the version of *qcr10* with intron 1 deleted and intron 2 retained, genomic DNA was amplified using two different primer pairs, as follows. The first primer pair used an upstream primer in the 5'UTR (primer #15 in Table 2.10) and a downstream chimeric primer whose 3' end aligned to the 3' end of exon 1, and whose 5' end aligned to the 5' end of exon 2 (primer #21). The second primer pair used an upstream chimeric primer whose 5' end aligned to the end of exon 1 and whose 3' end aligned to the start of exon 2 (primer #20), and a downstream primer that aligned to exon 3 (primers #16). The two PCR products were then gel purified and “stitched” together via the complementary ends produced by the chimeric primers in two cycles of PCR without external primers. Then the outermost primers – the upstream primer from the first primer pair (primer #15) and the downstream primer from the second primer pair (primer #16) – were added to an aliquot of the “stitched” template and standard PCR was carried out. Since the outermost

Table 2.10 – Primer sequences used

| Primer # | Gene | orientation | purpose | Figure | sequence |
|----------|----------------|-------------|---|-----------|---|
| 1 | <i>sim4</i> | F | Flanking PCR for detection of splicing | 2.1 | AAGACCTAACGAGGAGTTAAACAAG |
| 2 | <i>sim4</i> | R | Flanking PCR for detection of splicing | 2.1 | CTTTTGAGGGATCTTGGGGATC |
| 3 | <i>rpl4301</i> | F | Flanking PCR for detection of splicing | 2.1 | AAAGGTCGGCGTTACCGTAAATAT |
| 4 | <i>rpl4301</i> | R | Flanking PCR for detection of splicing | 2.1 | GAATGGTGGAGCGAGCACTAGTAGC |
| 5 | <i>srrm1</i> | F | Flanking PCR for detection of exon-skipping | 2.3 | GTTTATTGCTGCCTTCTTCATTCTT |
| 6 | <i>srrm1</i> | R | Flanking PCR for detection of exon-skipping | 2.3 | TGCGAGGCAGAGATAATTAACGA |
| 7 | <i>alp41</i> | F | Flanking PCR for detection of exon-skipping, and for construction of intron retained standards | 2.3, 2.S4 | GACATTGGGGGCGAGAAAAC |
| 8 | <i>alp41</i> | R | Flanking PCR for detection of exon-skipping; reverse primer for detection of both spliced isoforms by qPCR, and for construction of intron retained standards | 2.3, 2.S4 | TTTGATATTAAGGCCCGTTAATGC |
| 9 | <i>alp41</i> | F | junction-spanning primer between exons 4 and 5, for detection of canonically spliced isoform by qPCR | 2.3 | GTTGGTAGAAGAGAAACTTTTGTTTAC |
| 10 | <i>alp41</i> | F | junction-spanning primer between exons 4 and 6, for detection of exon-skipped isoform by qPCR | 2.3 | TGGTAGAAGAGATATTAATATTTCT |
| 11 | <i>alp41</i> | F | chimeric primer for creation of intron 4 - deleted, intron 5 - retained isoform of <i>alp41</i> | 2.S4 | ATTACAAGAATTGTTGGTAGAAGAGAACTTTTGTTTACTTCAATTTTGG |
| 12 | <i>alp41</i> | R | chimeric primer for creation of intron 4 - deleted, intron 5 - retained isoform of <i>alp41</i> | 2.S4 | CCAAAATTGAAGTAAACAAAAGTTTCTTCTACCAACAATCTTGTAAAT |
| 13 | <i>alp41</i> | F | chimeric primer for creation of intron 5 - deleted, intron 4 - retained isoform of <i>alp41</i> | 2.S4 | ACTTTCATCCGAAGAAATTAGCAAAATATTAATATTCTAAATATAAAT |
| 14 | <i>alp41</i> | R | chimeric primer for creation of intron 5 - deleted, intron 4 - retained isoform of <i>alp41</i> | 2.S4 | ATTTATATTAGAAATATTTAATATTTTGCTAATTTCTCGGATGAAAGT |
| 15 | <i>qcr10</i> | F | Flanking PCR for detection of splicing, and for construction of intron retained standards | 2.3, 2.S4 | AAAATAACTCCATCGACCTCCTCTC |
| 16 | <i>qcr10</i> | R | Flanking PCR for detection of splicing, and for construction of intron retained standards | 2.3, 2.S4 | CAAAGGCCCAACGGGAAAAG |
| 17 | <i>qcr10</i> | F | junction-spanning primer between exons 2 and 3, for detection of canonically spliced isoform by qPCR | 2.3 | GAACATGAAAACTATTCTGCCT |
| 18 | <i>qcr10</i> | F | junction-spanning primer between exons 1 and 3, for detection of exon-skipped isoform by qPCR | 2.3 | ATGATTTCTTTCTATTCTGCTC |
| 19 | <i>qcr10</i> | R | exon 5 reverse primer, for detection of both spliced isoforms by qPCR | 2.3 | CACCTGTCTTCTTCAGGAGTCTTGCT |
| 20 | <i>qcr10</i> | F | chimeric primer for creation of intron 1 - deleted, intron 2 - retained isoform of <i>qcr10</i> | 2.S4 | AAAATTAGTTACAATGATTTCTTTCTTTCCCAATAAGCCCATGTATCATG |
| 21 | <i>qcr10</i> | R | chimeric primer for creation of intron 1 - deleted, intron 2 - retained isoform of <i>qcr10</i> | 2.S4 | CATGATACATGGGCTTATTGGGAAAGAAAGAAATCATTTGAACATAATTTT |
| 22 | <i>qcr10</i> | F | chimeric primer for creation of intron 2 - deleted, intron 1 - retained isoform of <i>qcr10</i> | 2.S4 | TCACCCCTGAAAGAACTATGAAACTATTCTGCCTTTTCCCGTTGGGCC |
| 23 | <i>qcr10</i> | R | chimeric primer for creation of intron 2 - deleted, intron 1 - retained isoform of <i>qcr10</i> | 2.S4 | GGCCCAACGGGAAAAGGCAGGAATAGTTTTCATAGTTCTTTCAGGGGTG |

primers were the same as those used to carry out PCR of *qcr10* for Figure 2.3a, we could directly run out this stitched product with intron 1 deleted and intron 2 retained side by side with the cDNA PCR product from Figure 2.3a to test whether the band between the fully unspliced band and the fully spliced band migrated with the intron 2 - retained standard band. In an analogous fashion, a version of *qcr10* was constructed where intron 1 was retained and intron 2 was deleted. Similarly, versions of *alp41* were constructed where either intron 4 or intron 5 were retained. All of the primers used are indicated in Table 2.10. For each gene, both retained authentic standards were run on a gel alongside the heat shock (*alp41*) or cold shock (*qcr10*) cDNA PCR products from Figure 2.3a. The results are shown in Figure 2.S3, and are consistent with these bands representing amplicons with only one of the two introns excised.

Quantitative PCR for exon-skipping experiments of *alp41* and *qcr10*.

Quantitative PCR was used to examine the levels of the constitutively and alternatively spliced isoforms of both *alp41* and *qcr10*. Primers were designed to detect either the canonically spliced or the exon-skipped isoforms of each transcript, as follows. The forward primer in all cases was comprised entirely of sequence that spanned an exon-exon junction: the non-consecutive exons exon4-exon6 for the skipped isoform of *alp41* (primer #10 in Table 2.10), and the non-consecutive exons exon1-exon3 for the skipped isoform of *qcr10*

(primer #18 in Table 2.10); the consecutive exons exon4-exon5 for the canonical isoform of *alp41* (primer #9 in Table 2.10), and the consecutive exons exon2-exon3 for the canonical isoform of *qcr10* (primer #17 in Table 2.10). The number of bases in the primer corresponding to each exon was chosen to make the delta G as close to -17kcal/mol as possible, using a nearest neighbour calculation (64). The reverse primer in all cases aligned to a downstream exon: exon 6 for *alp41* (primer #8 in Table 2.10); and exon 5 for *qcr10* (primer #19 in Table 2.10). The standard curves for qPCR for both isoform-specific primer pairs were made using a dilution series of purified PCR product (see Figure 2.S4). Specifically, flanking PCR was performed for both *alp41* and *qcr10*, the products were run out on a gel, and the bands corresponding to the canonically and alternatively spliced isoforms were extracted from the gel. The gel-purified PCR products were then used to construct the standard curves shown in Figure 2.S4. The qPCR was performed in 15µL reactions containing 5ng cDNA, 250nM forward and reverse primers, 10mM TrisHCl (pH 8.5), 50mM KCl, 1.5mM MgCl₂, 0.2mM each dNTP, 0.25X Sybr Green, and 0.7ng of Taq DNA Polymerase. The amplification protocol was 40 cycles of 95°C for 15 seconds of denaturation, 56°C for 30 seconds of annealing and 72°C for 45 seconds of extension using a Roche LightCycler 480.

To induce heat shock, wild type *S. pombe* cells were grown to mid log phase (OD 0.6) in YES media at 25°C, then the culture was shifted to 37°C, after which cells were collected at 5, 10 and 20 minutes. A sample was also collected immediately prior to the temperature shift to serve as the baseline sample. To induce ethanol stress, wild type *S. pombe* cells were grown to mid log phase (OD 0.6) in YES media at 30°C, then were shifted to YES + 10% ethanol at 30°C. Cells were collected immediately prior to addition of ethanol (for baseline comparison), then after 5, 10 and 20 minutes of stress. Total RNA isolation and cDNA synthesis for each of these samples was performed as described above.

To induce cold shock, wild type *S. pombe* cells were grown overnight in YES at 30°C, then back-diluted and grown to mid log phase (OD 0.6) in YES media, then shifted to 16°C, after which cells were collected at 10, 30, 60, 120 and 180 minutes. A baseline sample was collected immediately prior to the temperature shift. To induce amino acid starvation, wild type *S. pombe* cells were grown to mid log phase (OD 0.6) in EMM at 30°C, then were shifted to EMM + 0.05M 3-amino-triazole (3-AT) at 30°C. Cells were collected immediately prior to addition of 3-AT (for baseline comparison), then after 10, 30, 60, 120 and 180 minutes of stress. Total RNA isolation and cDNA synthesis for each of these samples was performed as described above.

Testing specificity of isoform-specific qPCR primers for *alp41* and *qcr10*

For both of these genes, a qPCR experiment was performed to test the specificity of the two primer pairs used to detect the canonically spliced and the exon-skipped isoforms. For example, primers #10 and #8 from Table 2.10 were designed to specifically detect the exon-skipped isoform of *alp41* in the experiments shown in Figure 2.3b. To test the specificity of these primers for the alternative isoform of *alp41* (in this example termed the ‘cognate’ template) versus the canonically-spliced isoform of *alp41* (here termed the ‘non-cognate’ template), qPCR was carried out using these primers with a titration of increasing ratios of non-cognate to cognate template. The amount of cognate template was held constant at a level calculated to give the same threshold cycle as the lowest threshold cycle observed for this primer pair in the experiment in Figure 2.3b. The non-cognate template was doubled at each of eight steps, starting at a ratio of 1:4 (non-cognate: cognate) and ending at a ratio of 32:1, allowing for a >100-fold variation in the ratios. The source of the cognate and non-cognate template was gel-purified PCR products produced using the appropriate primers. The band corresponding to the canonical splicing event was purified and used as the non-cognate template in this case, and the band corresponding to the exon-skipping splicing event was purified and used as the cognate template. The test of specificity was carried out for all four primer

pairs used in the qPCR experiments shown in Figure 2.3b (*alp41* exon-skipping primer pair #10 and #9; *alp41* canonical splicing primer pair #9 and #8; *qcr10* exon-skipping primer pair #18 and #19; *qcr10* canonical splicing primer pair #17 and #19).

All qPCRs were performed in 15µL reactions containing templates as specified above, 250nM forward and reverse primers, 10mM TrisHCl (pH 8.5), 50mM KCl, 1.5mM MgCl₂, 0.2mM each dNTP, 0.25X Sybr Green, and 0.7ng of Taq DNA Polymerase. The amplification protocol for all *alp41* experiments was 40 cycles of 95°C for 15 seconds of denaturation, 56°C for 30 seconds of annealing and 72°C for 45 seconds of extension using a Roche LightCycler 480. The amplification protocol for all *qcr10* experiments was 40 cycles of 95°C for 15 seconds of denaturation, 55°C for 30 seconds of annealing and 72°C for 45 seconds of extension using a Roche LightCycler 480.

Importantly, as seen in Figure 2.S5a, high-specificity is observed for both of the primer pairs designed to detect the alternatively spliced isoforms. For the *alp41* primers, the presence of a 32-fold excess of the non-cognate template results in only a ~two-fold increase in the measured level of cognate material. Similarly, for the *qcr10* primers a 16-fold excess of the non-cognate template results in a two-fold increase in the measured level of cognate material. The primers designed to detect the constitutively spliced isoform of *qcr10* also

performed quite well, showing less than a two-fold change in the measured level of cognate template even in the presence of a 32-fold excess of non-cognate template. By contrast, the primers designed to detect the constitutively spliced isoform of *alp41* showed the highest level of cross-reactivity, demonstrating a two-fold change in the measured level of the cognate material in the presence of just a four-fold excess of the non-cognate template. It is important to note, however, that these data suggest that the precision with which the levels of the constitutive isoform of *qcr10* can be measured will be compromised under conditions where the alternative isoform is present in excess amounts. Because our data (and the current state of annotation for the *alp41* transcript) suggest that the alternative isoform is rare under normal conditions, we expect that these primers are giving a robust measure of the levels of the constitutively spliced isoform. Moreover, as seen in Figure 2.S5b, the amount of canonically spliced isoform for both *alp41* and *qcr10* that was detected in the experiments shown in Figure 2.3b is either decreasing or staying the same. As such, while the measured levels of alternative isoform at time zero might over-estimate their actual levels, the increased levels of alternative isoforms being detected over the course of the stress responses are highly unlikely to be derived from cross-reactivity with the constitutive isoforms.

Identifying gene and exon orthologs for skipped exons between *S. pombe* and *H. sapiens*

A perl script was written to find the reciprocal best BLAST (by blastp, default parameters) hit for every gene in the *S. pombe* genome. The set of genes and exons for *S. pombe* were defined by the pompep.fsa file at pombase and the gtf file from the BROAD Institute, and the set of genes and exons for humans was downloaded from ensembl, using the biomart tool (www.ensembl.org). For each *S. pombe* gene, the cDNA was translated and used to find a reciprocal best BLASTP hit with the human proteome. For every *S. pombe* gene that had a reciprocal best BLASTP hit (the human gene ortholog), every exon was subjected to a BLASTP search against every exon of the same length in the human gene ortholog. For those *S. pombe* – *H. sapiens* exon pairs of the same length, only if they were reciprocal best BLASTp hits and within 15 nucleotides in length were they considered exon orthologs.

Calculating the statistical significance of the conserved exon-skipping events.

We sought to address how likely it would be to observe at least two exons out of twenty-three randomly chosen *S. pombe* internal exons for which there was a

human orthologous exon that was subject to exon skipping, as follows. First, we calculated the proportion of all *S. pombe* internal coding exons (of which there are a total of 2641) for which there was a human ortholog by reciprocal best BLASTp as described above. A total of 113 such orthologs were identified when the exon lengths were required to be identical length, as was the case for the orthologous human exons in the *alp41* and *pth1* genes. Because the requirement for identical lengths could be considered too strict, we also identified a larger set of 263 orthologs whose lengths were within 15 nucleotides of each other (as described above). We then calculated the proportion of conserved exons that were simple cassette exons in humans, namely those conserved exons for which there was evidence that human orthologous exon could either be included or left out of a transcript produced from the host gene. This evidence was determined by using the ensembl human website (www.ensembl.org/Homo_sapiens/Info/Index), and manually inspecting every human gene for which there was an exon with an orthologous *S. pombe* exon for the presence of at least one transcript in which the exon in question was present and at least one in which it was absent. Absence here required that the exon was independent of flanking exons, such that it could be included or excluded in the transcript by itself, and not just as part of a larger exon skipping event involving multiple exons (i.e. we required that the exon was a simple cassette exon). All ensembl transcript evidence is based on EST or

protein data. This process of determining whether a conserved exon was also a simple cassette exon was the same process used to determine that the human orthologous exons to the *S. pombe* skipped *alp41* and *pth1* exons were also skipped in humans. Using this process, it was determined that for exon orthologs that were exactly the same length, the rate of skipping in humans was 13/113, and for exon orthologs that were within 15 nucleotides in length, the rate of skipping in humans was 40/263.

A perl script was written to determine how often one would expect to observe at least two exons that are conserved between *S. pombe* and humans and skipped in humans when picking 23 internal exons at random from the *S. pombe* genome. Specifically, a million trials were simulated whereby each trial consisted of twenty-three repeats of the following process. A number between 0 and 1 was randomly generated, then compared to the ratio of conservation and skipping (set to either 13/2641 or 40/2641, depending on the stringency set for conservation). If the random number was less than or equal to the conservation and skipping ratio, this was considered a “success”, and the count was incremented. After this process was repeated 23 times, the count was recorded. Each such set of 23 randomly generated numbers, comparisons and summations was considered a trial. Upon performing one million trials, the proportion of times that at least two “successes” were observed was reported

as the p-value of observing two conserved, skipped exons in our actual dataset. Using the more stringent conservation criterion of requiring exact length identity between orthologs (as was actually the case for our observed data with the conserved, skipped exons in the *alp41* and *pth1* genes) the p-value was 0.00567. Using the less stringent conservation criterion of requiring that exon orthologs be within 15 nucleotides of each other, the p-value was 0.0468. In both cases, the observed number of “successes” in the actual dataset is significantly higher than that expected by chance based on background conservation rates between *S. pombe* and human exons and background skipping rates in human exons that are conserved with *S. pombe* exons.

Finding exon orthologs for the *S. pombe* cassette exons in *alp41*, *srrm1*, *pth1* and *SPAC27D7.08c* among all NCBI eukaryotic sequenced genomes

A perl script was written to do the following. For each of the four *S. pombe* genes, the amino acid sequence of the full-length protein was used in a BLASTP against the entire non-redundant protein database from the NCBI:

(<http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins>), with a cutoff e-value of 5000 so as to not be too stringent at this step. Next, the reciprocal best BLASTP hit was calculated for each species with *S. pombe* to determine the set of orthologs at the gene level (the “gene ortholog” of the *S. pombe* gene). For

each gene ortholog, a search was conducted for the reciprocal best BLASTP hit among the translated exons of the *S. pombe* gene and the gene ortholog. Here the additional constraint of a length difference of at most 15 nucleotides was imposed. However, all exon orthologs in fact had identical lengths to the *S. pombe* exon.

Multiple sequence alignment

All non-BLAST multiple sequence alignments for Figures 2.4 and 2.S8 were performed using clustalw2, version 2.0.12, available from <http://www.ebi.ac.uk/Tools/msa/clustalw2/#>. The default options were used. For SPAC727D7.08c gaps in the alignment were suppressed.

Finding evidence for exon skipping of the *alp41*, *pth1*, and SPAC27D7.08c exons in macaque and mouse RNA-seq data.

In order to find evidence for exon-skipping of the four evolutionarily conserved exons in the *alp41*, *pth1*, SPAC27D7.08c and *srrm1* genes in species other than human and *S. pombe*, a recent RNA-seq study (65) that performed some of the deepest sequencing of several vertebrate species was searched. For each of the four *S. pombe* exons for which there was an orthologous exon in the species in question, a minimal unique alternative junction was created with which to search the RNA-seq data. An example of the process is outlined

below for the exon-skipping event in the *SPAC27D7.08c* gene in mouse. First, it was determined whether the *S. pombe* exon for this gene has an orthologous exon in mouse, as previously described. Since this is indeed the case, the next step was to determine the sequences of the exons flanking the conserved exon and to construct a set of junction sequences from these flanking exon sequences. To construct each junction sequence in the set, the last *n* bases of the upstream flanking exon were followed by the first *n* bases of the downstream flanking exon, where *n* was varied from 8 to 20. This produced a set of 13 alternative junction sequences with lengths from 16 to 40 nucleotides. To find the minimal unique junction (unique in the sense that it did not align anywhere to the mouse genome), all junctions were aligned using bowtie to the latest index of the mouse genome from the bowtie website. The shortest junction that did not align perfectly to the genome was considered the minimal unique junction. This process ensured that the junction used to search for alternative reads had at least 8 nucleotides from each of the exons flanking the conserved exon, and that it did not align anywhere in the genome (to reduce false positives). To search for reads indicating skipping of the conserved exon, the minimal unique junction was used in a unix grep command against the fastq file of read sequences. This command produced a list of lines in the read fastq file that contained the minimal unique junction. These lines were then manually curated to make sure that the junction (or “seed”) sequence could be extended

in at least one direction in the matched read, and that the extended sequence matched the flanking exon sequence. This process yielded several alternative splicing reads for macaque and one for mouse (confirming previous EST data). The procedure described above to search for alternative reads was also carried out to search for the corresponding canonical splicing reads, with the following modifications. Instead of constructing minimal unique junctions from the two exons flanking the conserved exon, this time two different pairs of exons were used: the upstream flanking exon and the conserved exon itself; and the conserved exon and the downstream flanking exon. Whereas for the alternative junction construction two non-contiguous exons $n-1$ and $n+1$ were used, in this case the two pairs $n-1, n$ and $n, n+1$ were used to construct the canonical junctions. The read counts for macaque and mouse for all four genes are given in Table 2.11.

Table 2.11 – Exon skipping reads for Macaque and Mouse from Burge *et al.*

| organism id | organism name | NCBI SRA Run ID | Total Reads | Avg Read Length |
|-------------|---------------|-----------------|-------------|-----------------|
| 9544 | Macaque | SRR594455 | 215339102 | 80 |
| 9544 | Macaque | SRR594456 | 209648898 | 80 |
| 9544 | Macaque | SRR594458 | 217275344 | 80 |
| 9544 | Macaque | SRR594460 | 225465734 | 80 |
| 9544 | Macaque | SRR594461 | 229927714 | 80 |
| 10090 | Mouse | SRR594405 | 268091442 | 80 |

| organism id | organism name | alp41 alternative | alp41 upstream constitutive | alp41 downstream constitutive | alp41 avg constitutive |
|-------------|---------------|-------------------|-----------------------------|-------------------------------|------------------------|
| 9544 | Macaque | 1 | 413 | 728 | 570.5 |
| 9544 | Macaque | 0 | 140 | 179 | 159.5 |
| 9544 | Macaque | 0 | 188 | 404 | 296 |
| 9544 | Macaque | 1 | 129 | 210 | 169.5 |
| 9544 | Macaque | 0 | 94 | 188 | 141 |
| 10090 | Mouse | 0 | 17 | 31 | 24 |

| organism id | organism name | pth1 alternative | pth1 upstream constitutive | pth1 downstream constitutive | pth1 avg constitutive |
|-------------|---------------|------------------|----------------------------|------------------------------|-----------------------|
| 9544 | Macaque | 1 | 10 | 40 | 25 |
| 9544 | Macaque | 1 | 23 | 43 | 33 |
| 9544 | Macaque | 1 | 42 | 105 | 73.5 |
| 9544 | Macaque | 1 | 20 | 66 | 43 |
| 9544 | Macaque | 1 | 6 | 33 | 19.5 |
| 10090 | Mouse | 0 | 12 | 12 | 12 |

| organism id | organism name | SPAC27D7.08c alternative | SPAC27D7.08c upstream constitutive | SPAC27D7.08c downstream constitutive | SPAC27D7.08c avg constitutive |
|-------------|---------------|--------------------------|------------------------------------|--------------------------------------|-------------------------------|
| 9544 | Macaque | n/a | n/a | n/a | n/a |
| 9544 | Macaque | n/a | n/a | n/a | n/a |
| 9544 | Macaque | n/a | n/a | n/a | n/a |
| 9544 | Macaque | n/a | n/a | n/a | n/a |
| 9544 | Macaque | n/a | n/a | n/a | n/a |
| 10090 | Mouse | 1 | 38 | 17 | 27.5 |

*n/a indicates that there is no orthologous exon in this species for this gene

Assessing Tissue-specific Exon skipping of Human ARL2

Using Hi-Seq read data from the Illumina Human Body Map 2.0 Project (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30611>), reads that corresponded to the junction of ARL2 exon4–exon5 or exon 5–exon6 (canonical splicing reads) and reads that corresponded to the junction of exon4–exon6 (alternative splicing reads) were tabulated for each of sixteen different human tissues (Table 2.12). For each tissue, a percentage of exon skipping was calculated by dividing the exon 5 skipping reads by the number of canonical splicing reads divided by two.

The data were tested for heterogeneity in exon5 skipping percentage using the Chi-square test in the R statistical package on the 16 x 2 table of tissue-specific alternative and canonical junction read counts. The p-value for this test was 3.77×10^{-5} . By instead comparing the 2 x 2 table of canonical and alternative splicing read counts from the tissue with the highest percentage of exon 5 skipping (lung, 19.15%) to the counts from the tissue with the lowest percentage (skeletal muscle, 1.69%) by Fisher's exact test, the p-value was 4.48×10^{-5} . The analogous test comparing the tissues with second-highest and second-lowest exon5-skipping yielded a p-value of 2.46×10^{-4} .

Table 2.12 – Tissue specific ARL2 exon skipping

| Tissue | skipped reads | canonical reads | 100% * skipped / (canonical / 2) |
|-----------------|----------------------|------------------------|---|
| adipose | 12 | 304 | 7.894736842 |
| adrenal | 10 | 279 | 7.168458781 |
| brain | 7 | 281 | 4.982206406 |
| breast | 9 | 286 | 6.293706294 |
| colon | 8 | 283 | 5.653710247 |
| heart | 7 | 198 | 7.070707071 |
| kidney | 4 | 237 | 3.375527426 |
| liver | 1 | 63 | 3.174603175 |
| lung | 18 | 188 | 19.14893617 |
| lymph node | 8 | 225 | 7.111111111 |
| ovary | 9 | 269 | 6.691449814 |
| prostate | 18 | 255 | 14.11764706 |
| skeletal muscle | 2 | 236 | 1.694915254 |
| testes | 3 | 306 | 1.960784314 |
| thyroid | 5 | 320 | 3.125 |
| white blood | 7 | 179 | 7.82122905 |

References

1. Black DL (2000) Protein Diversity from Alternative Splicing: A Challenge for Bioinformatics and Post-Genome Biology. *Cell* 103:367–370.
2. Nilsen TW, Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463:457–463.
3. Lareau LF, Brooks AN, Soergel DAW, Meng Q, Brenner SE (2007) The coupling of alternative splicing and nonsense-mediated mRNA decay. *Adv Exp Med Biol* 623:190–211.
4. McGlincy NJ, Smith CWJ (2008) Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends in Biochemical Sciences* 33:385–393.
5. Breitbart RE, Andreadis A, Nadal-Ginard B (1987) Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annu Rev Biochem* 56:467–495.
6. Kim E, Magen A, Ast G (2007) Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res* 35:125–131.
7. Fedorov A, Merican AF, Gilbert W (2002) Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proceedings of the National Academy of Sciences* 99:16128–16133.
8. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol* 13:1512–1517.
9. Csurös M, Rogozin IB, Koonin EV (2008) Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum-likelihood approach. *Mol Biol Evol* 25:903–911.
10. Schwartz SH et al. (2008) Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res* 18:88–103.
11. Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* 11:345–355.

12. Csuros M, Rogozin IB, Koonin EV (2011) A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol* 7:e1002150.
13. Irimia M, Rukov JL, Penny D, Roy SW (2007) Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. *BMC Evol Biol* 7:188.
14. Roy SW, Irimia M (2009) Splicing in the eukaryotic ancestor: form, function and dysfunction. *Trends Ecol Evol (Amst)* 24:447–455.
15. Sorber K, Dimon MT, DeRisi JL (2011) RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. *Nucleic Acids Res* 39:3820–3835.
16. Loftus BJ et al. (2005) The Genome of the Basidiomycetous Yeast and Human Pathogen *Cryptococcus neoformans*. *Science* 307:1321–1324.
17. Yu B et al. (2011) Spliceosomal genes in the *D. discoideum* genome: a comparison with those in *H. sapiens*, *D. melanogaster*, *A. thaliana* and *S. cerevisiae*. *Protein & Cell* 2:395–409.
18. Wood V et al. (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature* 415:871–880.
19. Bitton DA et al. (2011) Augmented annotation of the *Schizosaccharomyces pombe* genome reveals additional genes required for growth and viability. *Genetics* 187:1207–1217.
20. Prabhala G, Rosenberg GH, Käufer NF (1992) Architectural features of pre-mRNA introns in the fission yeast *Schizosaccharomyces pombe*. *Yeast* 8:171–182.
21. Kupfer DM et al. (2004) Introns and splicing elements of five diverse fungi. *Eukaryotic Cell* 3:1088–1100.
22. Lützelberger M, Gross T, Käufer NF (1999) Srp2, an SR protein family member of fission yeast: in vivo characterization of its modular domains. *Nucleic Acids Res* 27:2618–2626.
23. Tang Z, Käufer NF, Lin R-J (2002) Interactions between two fission yeast serine/arginine-rich proteins and their modulation by phosphorylation. *Biochem J* 368:527–534.

24. Webb CJ, Romfo CM, Van Heeckeren WJ, Wise JA (2005) Exonic splicing enhancers in fission yeast: functional conservation demonstrates an early evolutionary origin. *Genes Dev* 19:242–254.
25. Sarmah B, Chakraborty N, Chakraborty S, Datta A (2002) Plant pre-mRNA splicing in fission yeast, *Schizosaccharomyces pombe*. *Biochem Biophys Res Commun* 293:1209–1216.
26. Käufer NF, Simanis V, Nurse P (1985) Fission yeast *Schizosaccharomyces pombe* correctly excises a mammalian RNA transcript intervening sequence. *Nature* 318:78–80.
27. Kishida M, Nagai T, Nakaseko Y, Shimoda C (1994) Meiosis-dependent mRNA splicing of the fission yeast *Schizosaccharomyces pombe* *mes1+* gene. *Curr Genet* 25:497–503.
28. Moldón A et al. (2008) Promoter-driven splicing regulation in fission yeast. *Nature* 455:997–1000.
29. Wilhelm BT et al. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453:1239–1243.
30. Rhind N et al. (2011) Comparative functional genomics of the fission yeasts. *Science* 332:930–936.
31. Ruskin B, Green MR (1985) An RNA processing activity that debranches RNA lariats. *Science* 229:135–140.
32. Domdey H et al. (1984) Lariat structures are in vivo intermediates in yeast pre-mRNA splicing. *Cell* 39:611–621.
33. Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C (2007) Rapid, transcript-specific changes in splicing in response to environmental stress. *Mol Cell* 27:928–937.
34. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
35. Lim LP, Burge CB (2001) A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci USA* 98:11193–11198.
36. Bicknell AA, Cenik C, Chua HN, Roth FP, Moore MJ (2012) Introns in UTRs: why we should stop ignoring them. *Bioessays* 34:1025–1034.

37. Isken O, Maquat LE (2008) The multiple lives of NMD factors: balancing roles in gene and genome regulation. *Nat Rev Genet* 9:699–712.
38. Leeds P, Peltz SW, Jacobson A, Culbertson MR (1991) The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing a premature translational termination codon. *Genes & Development* 5:2303–2314.
39. Brooks AN et al. (2011) Conservation of an RNA regulatory map between *Drosophila* and mammals. *Genome Res* 21:193–202.
40. Wang K et al. (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* 38:e178.
41. Kalyna M, Lopato S, Voronin V, Barta A (2006) Evolutionary conservation and regulation of particular alternative splicing events in plant SR proteins. *Nucleic Acids Res* 34:4395–4405.
42. Iida K, Go M (2006) Survey of conserved alternative splicing events of mRNAs encoding SR proteins in land plants. *Mol Biol Evol* 23:1085–1094.
43. Brandenberger R et al. (2004) Transcriptome characterization elucidates signaling networks that control human ES cell growth and differentiation. *Nat Biotechnol* 22:707–716.
44. Strausberg RL et al. (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci USA* 99:16899–16903.
45. Merkin J, Russell C, Chen P, Burge CB (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* 338:1593–1599.
46. Carninci P et al. (2005) The transcriptional landscape of the mammalian genome. *Science* 309:1559–1563.
47. Rogozin IB, Carmel L, Csuros M, Koonin EV (2012) Origin and evolution of spliceosomal introns. *Biol Direct* 7:11.
48. Cheng C, Sharp PA (2006) Regulation of CD44 alternative splicing by SRm160 and its potential role in tumor cell invasion. *Mol Cell Biol* 26:362–370.
49. Szymczyna BR et al. (2003) Structure and function of the PWI motif: a novel nucleic acid-binding domain that facilitates pre-mRNA processing. *Genes Dev* 17:461–475.

50. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* 446:926–929.
51. Ni JZ et al. (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev* 21:708–718.
52. Shao W, Kim H-S, Cao Y, Xu Y-Z, Query CC (2012) A U1-U2 snRNP interaction network during intron definition. *Mol Cell Biol* 32:470–478.
53. Fox-Walsh KL et al. (2005) The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci USA* 102:16176–16181.
54. Sharma S, Kohlstaedt LA, Damianov A, Rio DC, Black DL (2008) Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nat Struct Mol Biol* 15:183–191.
55. Clement JQ, Qian L, Kaplinsky N, Wilkinson MF (1999) The stability and fate of a spliced intron from vertebrate cells. *RNA* 5:206–220.
56. Grabherr MG et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644–652.
57. Oshlack A, Wakefield MJ (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* 4:14.
58. Taggart AJ, DeSimone AM, Shih JS, Filloux ME, Fairbrother WG (2012) Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat Struct Mol Biol* 19:719–721.
59. Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* 14:1188–1190.
60. Letunic I, Bork P (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* 39:W475–478.
61. Forsburg SL, Rhind N (2006) Basic methods for fission yeast. *Yeast* 23:173–183.
62. Quail MA et al. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5:1005–1010.

63. Lim LP, Burge CB (2001) A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci USA* 98:11193–11198.
64. Sugimoto N, Nakano S, Yoneyama M, Honda K (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res* 24:4501–4505.
65. Merkin J, Russell C, Chen P, Burge CB (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* 338:1593–1599.
66. Robinson JT et al. (2011) Integrative genomics viewer. *Nature Biotechnology* 29:24–26.

CHAPTER 3: OPTIMIZING LARIAT SEQUENCING TO BETTER PROFILE THE WHOLE *S. POMBE* INTRONOME

3.1 Introduction

In the first lariat sequencing experiment described in Chapter 2, scores of new introns were discovered, as well as intron retention events and the first examples of exon-skipping known in the unicellular yeast, *Schizosaccharomyces pombe*. However, since this experiment was optimized for profiling lariats with sizes corresponding to single-exon skipping event with lengths between 85 and 200nt (see Chapter 2), and since these are larger than most introns in *S. pombe*, which have a modal length below 50 nucleotides (Wood et al., 2002, 2012), the first lariat experiment left a significant portion of the *S. pombe* intronome unprofiled.

During sequence library preparation for the approach used in Chapter 2, there were two major steps that were reasoned to contribute to the loss of shorter lariats below 85 nucleotides (Figure 3.1b). The first of these was random-priming of lariats for cDNA synthesis: the shortest lariats are calculated to have a length under 30 nucleotides based on intron lengths, and it is possible that the random nine-mers used to prime reverse transcriptase did not do so efficiently for these small lariat species. The second potential cause of

loss of small lariats is the column purification step used in the method after the second strand cDNA synthesis. Since the double stranded cDNA produced by random priming of lariats is at most as long as the original intron producing the lariat, except in the relatively rare cases of branchpoint read-through by Reverse Transcriptase (Pratico & Silverman, 2007), it is possible that lariats with lengths shorter than 85 nucleotides were depleted at this step. To address these issues, I modified the lariat sequencing library preparation protocol to first debranch lariats and add RNA primers prior to cDNA synthesis, instead of performing random-primed cDNA synthesis on intact lariat RNA. Thus, cDNA synthesis was performed on linear molecules (not potentially small circular lariats), and there was no column purification prior to adding the adapters, which lengthened the linearized lariats.

3.2 Results

RNA was isolated from $\Delta dbr1$ *S. pombe* grown under a variety of conditions and stresses (see Methods) to elicit a wide variety of transcriptional and splicing programs. In contrast with the previous approach, RNA from cells grown under normal rich media conditions (RNA from the “rich” sample -- see Methods) was separated from a pool of RNA isolated from cells grown under each of the stress conditions (RNA from the “stress” sample -- see Methods), and these samples were kept apart for the entirety of the sequencing library preparation, receiving different sequencing barcodes to allow them to be distinguished. For both the “rich” and the “stress” RNA samples, two-dimensional polyacrylamide gel electrophoresis was carried out using two different sets of gel conditions optimized to isolate shorter and longer lariats (see Methods). This approach resulted in four lariat RNA samples eluted from the lariat gel arcs (Figure 3.1a) – “Short Rich”, “Short Stress”, “Long Rich” and “Long Stress”.

To avoid depletion of shorter lariats and better profile the *S. pombe* intronome, the sequence library approach used in Chapter 2 was adapted in two main ways (Figure 3.1b). First, prior to cDNA synthesis, lariats were linearized using the debranchase (DBR1) enzyme (Khalid, Damha, Shuman, & Schwer, 2005; Zhang, Hesselberth, & Fields, 2007), and then the linearized

lariats were fragmented. After RNA fragmentation, the second major modification was adapter ligation to the fragmented linearized lariats prior to adapter-primer cDNA synthesis. This approach ensured that both potential sources of depletion of short lariats were removed: cDNA synthesis does not rely on priming within short circular RNAs, and adapters are ligated prior cDNA synthesis, and column-based cleanup is replaced by precipitation (see Methods).

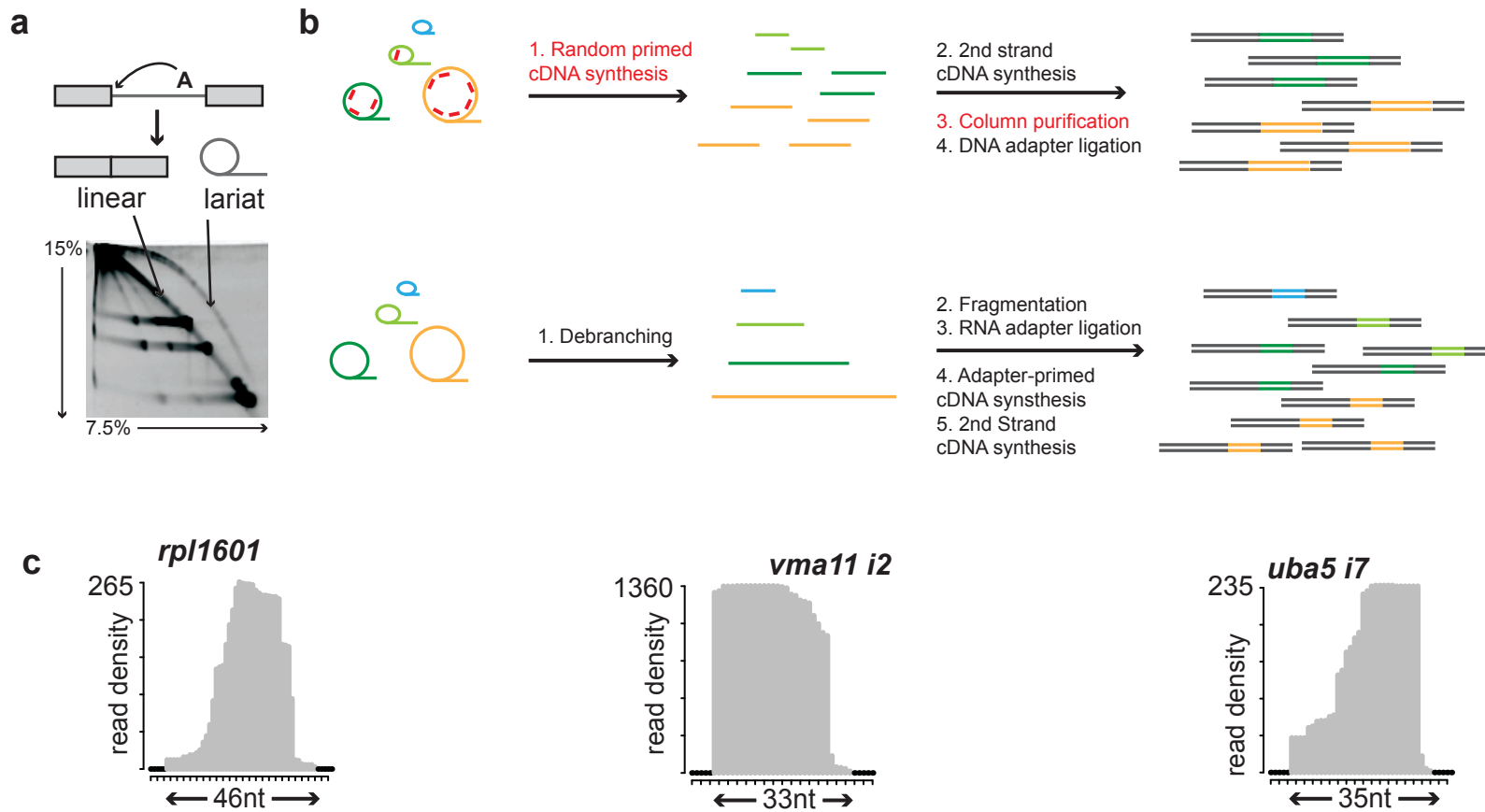


Figure 3.1

The revised lariat sequencing method is designed to capture shorter lariats. **(a)** Two-dimensional gel electrophoresis to separate lariats from linear mRNA is carried out for both methods. **(b)** Schematic overview of previous lariat sequencing library preparation (top) and revised preparation (bottom). Lariat species of different sizes are indicated by different colours, with the longest in cream, then dark green, then light green, then light blue. Red line segments depict random 9-mers used for RT priming. Library adapter oligos are indicated in dark gray. Arrows indicate steps of library production, with red text highlighting procedures that potentially deplete shorter lariats. **(c)** Read density plots (from the “Short Rich” sample) for three very short introns (two of which are in the shortest 10 introns in the *S. pombe* genome) that are not detected in the previous lariat sequencing dataset, but are well detected in the revised version.

Table 3.1

Alignable read counts for each sample

| Sample | alignable reads |
|--------------|-----------------|
| | |
| Short Rich | 2,366,974 |
| Short Stress | 1,334,222 |
| Long Rich | 5,006,071 |
| Long Stress | 14,202,325 |

The Modified Lariat Sequencing Approach (LarSeq2) Captures Short Introns

Upon trimming adapters sequences and aligning reads to the *S. pombe* genome (see Methods and Table 3.1), the dataset was tested for recovery of short introns as follows. Every non-zero read region (“peak”) with at least five reads was assayed for overlap with an annotated *S. pombe* intron. Those that did so with a maximum of five nucleotides between the 5’ ends of the peak and the intron and 10 nucleotides between the 3’ ends were considered as “intronic peaks” (see Methods). Read density plots for four of the shortest of these – none of which was detected in the previous lariat sequencing method – are shown in Figure 3.1c.

To test more globally for recovery of short introns, a distribution of the lengths of these intronic peaks was plotted along with the distribution of the lengths of all annotated introns in the *S. pombe* genome (Figure 3.2a) (Wood et al., 2002, 2012). For comparison, an analogous plot was made for introns profiled by the previous lariat sequencing method (Figure 3.2b, see Chapter 2, methods). As for longer introns, the new approach was compared to the old one by counting the number of previously annotated introns with a length greater than 200 nucleotides detected with at least 25 reads by each method (compare Table 2.4 for LarSeq1 to Table 3.3 for LarSeq2). The new method detects 172 such longer introns, whereas the old method detects 93. Thus the

new method is strikingly better than the old method at profiling the entire range of *S. pombe* introns. To more closely examine the breakdown of profiled intron lengths by sample, density plots were made comparing the intron length distributions of the Short Rich and Long Rich samples, and those of the Short Stress and Long Stress samples (Figure 3.3).

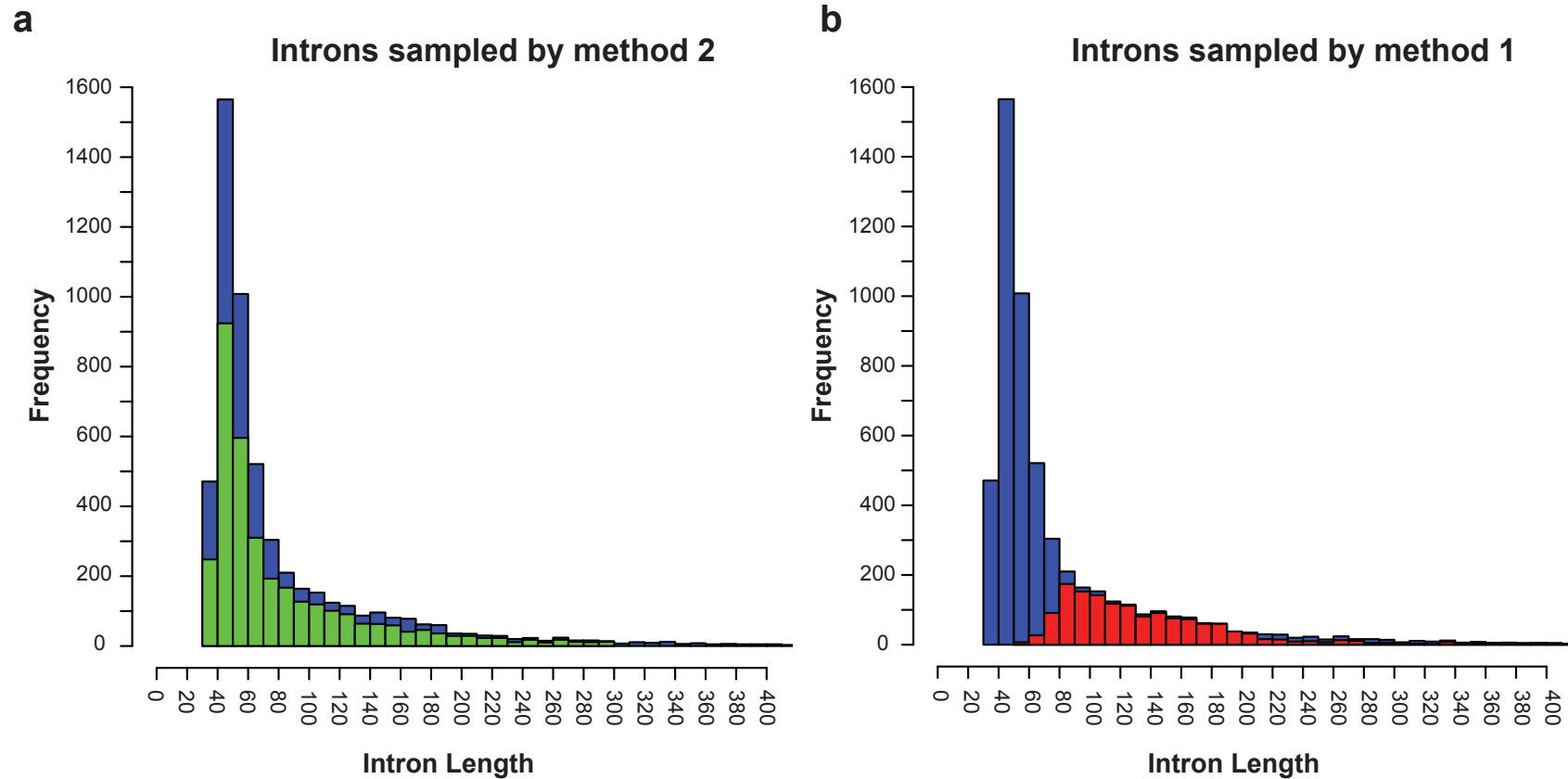


Figure 3.2

Length distributions for introns detected by both lariat sequencing methods. **(a)** Length distributions of introns detected by the revised lariat sequencing method (green), overlaid on the length distribution of all annotated *S. pombe* introns (blue). **(b)** Length distributions of introns detected by the previous lariat sequencing method (red), overlaid on the length distribution of all annotated *S. pombe* introns (blue).

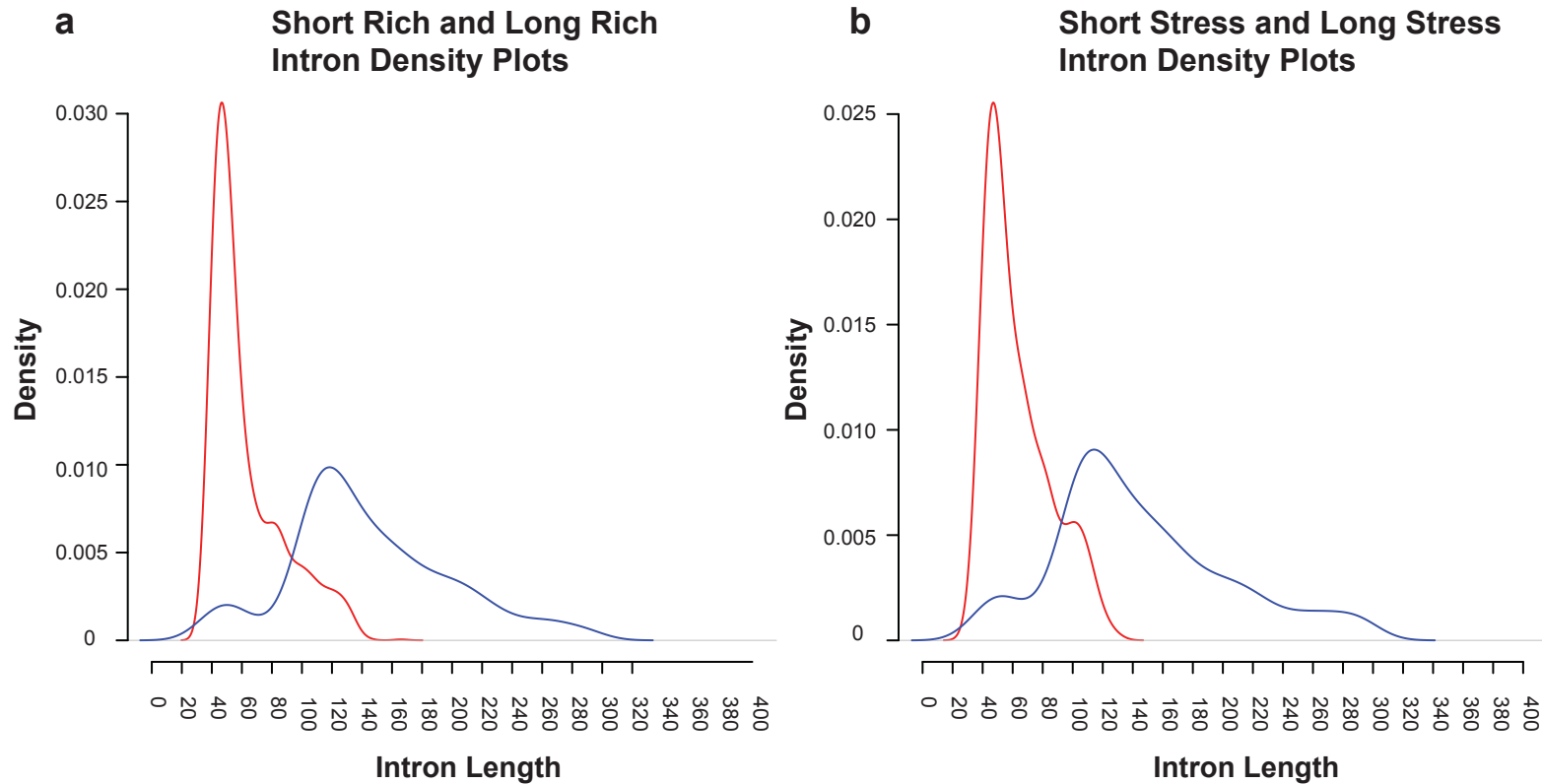


Figure 3.3

Density plots of lengths of introns detected in the revised lariat sequencing method. **(a)** Density plots of length of introns detected in the Short Rich (red) and Long Rich (blue) samples. **(b)** Density plots of length of introns detected in the Short Stress (red) and Long Stress (blue) samples.

LarSeq2 confirms over 70% of the from the previous study and Discovers Hundreds of Novel Introns

For each of the four samples, the boundaries of peaks with over twenty-five reads were searched for five prime splice sites, branch point sequences and three prime splice sites as described previously (see Chapter 2 and Methods). Peaks for which such sequences were found (“inferred” introns) were scored and classified (see Methods) into three main categories with several subcategories each (Table 3.2 and Table 3.3*). The largest category, with 3395 inferred introns, comprised introns in the *S. pombe* genome previously annotated independently of lariat sequencing (Wood et al., 2002, 2012). These are referred to as “known” introns. A second category corresponded to introns discovered in the previous lariat sequencing experiment (Chapter 2), referred to as “larseq1” introns. There were 164 of these, meaning that over 70% of the 222 introns inferred previously are confirmed with an independent biological experiment. The final category is completely novel introns, of which LarSeq2 discovers 555. 175 of these occur in previously intronless genes, which represents almost 3% of the annotated genes in the genome. Of the 100 novel introns discovered in noncoding RNA, 20 are discovered in intergenic lncRNAs.

Table 3.2

Summary of detected introns across all four samples by category. “known” refers to introns annotated independently of lariat sequencing. “larseq1” refers to introns not in the “known” category, but detected in the previous lariat sequencing method. “novel” refers to intron in neither the “known” nor the “larseq1” categories, but detected with the revised lariat sequencing method. “prev intronless” refers to novel introns that occur in a gene that was previously annotated as having no introns. See Methods for description of all categories.

| class | count |
|--------------------------------------|--------------|
| | |
| known_intron_alt_3ss | 305 |
| known_intron_alt_5ss | 290 |
| known_single_intron | 2710 |
| known_single_intron_alt_splice_sites | 13 |
| larseq1_intron_alt_3ss | 33 |
| larseq1_intron_alt_5ss | 42 |
| larseq1_intron_alt_splice_sites | 2 |
| larseq1_single_intron | 79 |
| exon_skip | 47 |
| multi | 120 |
| novel_3utr_intron | 65 |
| novel_5utr_intron | 39 |
| novel_antisense | 102 |
| novel_cds_frameshift | 89 |
| novel_cds_in-frame | 39 |
| novel_unclassified | 46 |

| | |
|--|------|
| prev_intronless_novel_3utr_intron | 18 |
| prev_intronless_novel_5utr_intron | 26 |
| prev_intronless_novel_cds_frameshift | 38 |
| prev_intronless_novel_cds_in-frame | 11 |
| prev_intronless_novel_ncRNA | 82 |
| | |
| total known | 3318 |
| total larseq1 | 156 |
| total novel in genes with annotated introns | 380 |
| total novel in previously intronless genes | 175 |
| exon skip | 47 |
| multi | 82 |
| total | 4158 |
| | |
| novel introns with RNA-seq junction support | 92 |

Since LarSeq2 uses a strand-specific sequencing approach (see Methods), it is possible to unequivocally assign certain introns as occurring in RNA antisense to the annotated coding regions of genes. Two examples overlapping exons of the *prp22* and *rds1* genes were shown to occur on the antisense to these genes as predicted, using strand-specific RT-PCR (Figure 3.4).

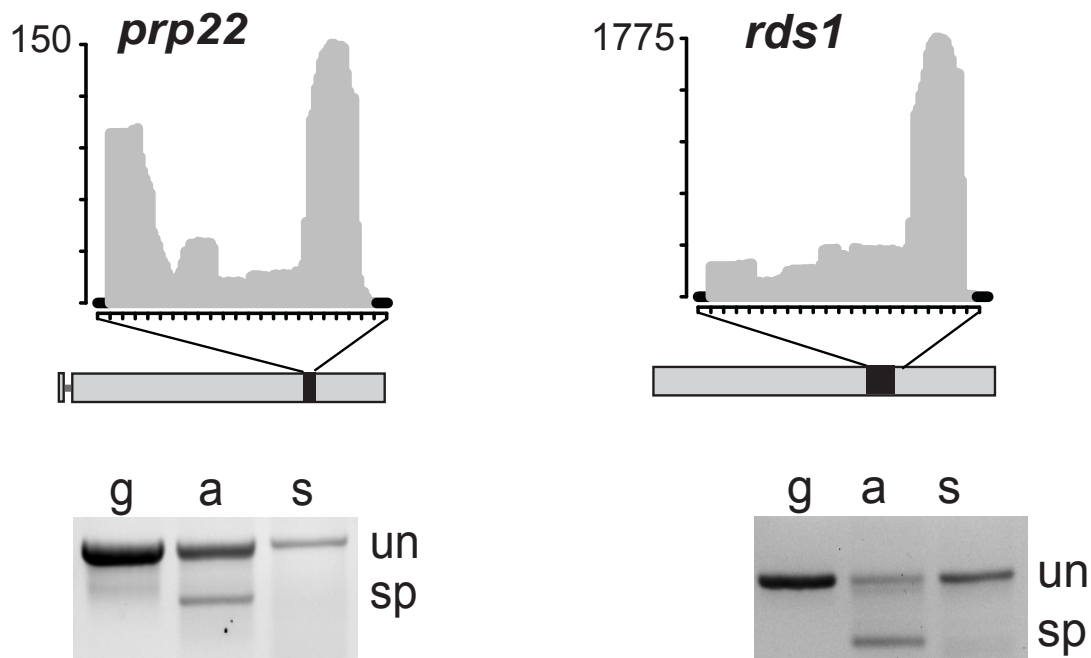


Figure 3.4

Validation of splicing of introns that occur antisense to annotated protein coding genes. Top: read density plots for the introns antisense to the *prp22* and *rds1* genes in the “Short Rich” samples. Gene structure is shown as a cartoon underneath, with zoomed in regions indicated using black rectangles and lines. Bottom: products of flanking PCR (using primers flanking the peaks) with three different templates. g = “genomic DNA”, “a” = cDNA from RNA transcribed from the strand antisense to the protein coding gene, “s” = cDNA from RNA transcribed from the strand harbouring the protein coding gene. “un” = unspliced isoform, “sp” = spliced isoform.

Scores of Introns Show Differential Expression in the Rich vs. Stress Samples

To assess differential expression of introns, normalized read counts were compared between the Short Rich and Short Stress samples (the “short” comparison, as well as between the Long Rich and Long Stress samples (the “long” comparison, see Methods, Table 3.4*). Log2 values of the ratio of normalized “stress” reads to normalized “rich” reads were calculated for each comparison, with the top twenty most up and down-regulated introns shown in tables 3.5 – 3.8. To test whether differential expression of introns was correlated with differential expression of genes, the log2 values for stress vs. rich intron expression were compared to log2 values for stress vs. rich gene expression from a previous RNA-seq experiment (See methods) (Rhind et al., 2011). In all of the top differentially expressed introns, most genes are also differentially expressed in the same way, suggesting a potential stress-related modulation of transcription of the genes harbouring these introns, rather than a specific change in splicing efficiency for these introns. For several of the introns however, host gene expression is relatively stable between rich and stress conditions (Tables 3.5 – 3.8, rows highlighted in red), suggesting potential stress-induced changes in splicing efficiency.

| gene | intron | intron length | gene short name | gene description | log2 intron stress/rich | log2 gene rpkm stress / rich |
|---------------|---------|---------------|-----------------|---|-------------------------|------------------------------|
| SPBC15C4.02 | intron2 | 92 | SPBC15C4.02 | ABC1 kinase family protein | 2.65283 | -0.052 |
| SPAC1039.10 | intron2 | 56 | mmf2 | homologous Pmf1 factor 1, implicated in isoleucine biosynthesis | 2.4501 | 0.175 |
| SPBC17F3.02 | intron2 | 65 | nak1 | PAK-related kinase Nak1 | 2.44106 | 0.926 |
| SPCC11E10.07c | intron3 | 96 | SPCC11E10.07c | translation initiation factor eIF2B alpha subunit | 2.36667 | -0.215 |
| SPAC23H4.17c | intron1 | 65 | srb10 | cyclin-dependent protein Srb mediator subunit kinase Srb10 | 2.23182 | 0.051 |
| SPBC12D12.08c | intron2 | 71 | ned8 | ubiquitin-like protein modifier Ned8 | 2.21434 | 0.191 |
| SPBC1604.04 | intron2 | 93 | SPBC1604.04 | mitochondrial thiamine pyrophosphate transporter | 2.21006 | -0.023 |
| SPAC17A5.07c | intron2 | 58 | ulp2 | SUMO deconjugating cysteine peptidase Ulp2 | 2.0784 | -0.24 |
| SPAC13G7.10 | intron1 | 87 | mug152 | Myb family telomere binding protein | 1.99563 | 0.022 |
| SPAC19A8.16 | intron1 | 111 | prl65 | conserved fungal protein | 1.93827 | 5.204 |
| SPAC11E3.06 | intron1 | 84 | map1 | MADS-box transcription factor Map1 | 1.92344 | 0.42 |
| SPAC3A11.13 | intron2 | 78 | SPAC3A11.13 | prefoldin subunit 6 | 1.86541 | 0.049 |
| SPCC622.14 | intron2 | 78 | SPCC622.14 | GTPase activating protein | 1.82215 | -0.688 |
| SPAC139.06 | intron4 | 71 | hat1 | histone acetyltransferase Hat1 | 1.76015 | -0.382 |
| SPAC1F3.09 | intron4 | 70 | mug161 | Cwfj family protein, splicing factor | 1.75763 | 1.511 |
| SPCC790.02 | intron2 | 50 | pep3 | HOPS/CORVET complex subunit, ubiquitin-protein ligase E3 | 1.72262 | 0.222 |
| SPAC3G6.11 | intron2 | 59 | chl1 | ATP-dependent DNA helicase Chl1 | 1.71148 | -0.123 |

| | | | | | | |
|--------------|---------|----|--------------|---|---------|-------|
| SPCC1494.09c | intron1 | 68 | SPCC1494.09c | sequence orphan | 1.6769 | 0.642 |
| SPBC691.01 | intron4 | 61 | pfa5 | vacuolar membrane palmitoyltransferase Pfa5 | 1.65735 | 0.09 |
| SPAC23D3.07 | intron1 | 65 | pup1 | 20S proteasome complex subunit beta 2 | 1.65642 | 0.234 |

Table 3.5

Top twenty up-regulated introns in the Long Stress sample vs. the Long Rich sample.

| gene | intron | intron length | gene short name | gene description | log2 intron stress/rich | log2 gene rpkm stress / rich |
|--------------|----------|---------------|-----------------|--|-------------------------|------------------------------|
| SPBC1921.01c | intron1 | 131 | rpl35b | 60S ribosomal protein L35a | -6.89923 | -1.381 |
| SPBC18H10.02 | intron1 | 126 | lcf1 | long-chain-fatty-acid-CoA ligase Lcf1 | -6.83794 | -1.22 |
| SPCC777.12c | intron1 | 120 | SPCC777.12c | thioredoxin family protein | -5.04917 | -0.043 |
| SPAC732.01 | intron2 | 32 | vma11 | V-type ATPase V0 proteolipid subunit | -4.83141 | -0.214 |
| SPAC4F10.20 | intron3 | 123 | grx1 | glutaredoxin Grx1 | -3.93228 | 0.658 |
| SPAC29A4.08c | intron1 | 40 | prp19 | ubiquitin-protein ligase E4 | -3.7642 | -0.401 |
| SPAC144.05 | intron2 | 40 | SPAC144.05 | ATP-dependent DNA helicase | -3.68866 | -0.368 |
| SPBC16D10.02 | intron1 | 43 | trm11 | tRNA (guanine-N2-)-methyltransferase catalytic subunit | -3.51028 | -1.239 |
| SPBC1734.07c | intron1 | 40 | SPBC1734.07c | TRAPP complex subunit Trs85 | -3.43282 | -0.095 |
| SPAPB21F2.03 | intron1 | 41 | slx9 | ribosome biogenesis protein Slx9 | -3.42067 | -0.669 |
| SPAC13G7.08c | intron2 | 43 | crb3 | WD repeat protein Crb3 | -3.39803 | -1.289 |
| SPAC17D4.01 | intron2 | 129 | pex7 | peroxin-7 | -3.35756 | -0.525 |
| SPAC17C9.05c | intron4 | 49 | pmc3 | mediator complex subunit Pmc3/Med27 | -3.33539 | -0.052 |
| SPBC8D2.13 | intron3 | 45 | shq1 | box H/ACA snoRNP assembly protein Shq1 | -3.31459 | -0.588 |
| SPAC1D4.05c | intron1 | 52 | SPAC1D4.05c | Erd1 homolog | -3.29646 | -0.063 |
| SPCC622.19 | intron1 | 43 | jmj4 | Jmj4 protein | -3.29435 | -1.183 |
| SPBC29A3.14c | intron15 | 42 | trt1 | telomerase reverse transcriptase 1 protein Trt1 | -3.29305 | -0.241 |

| | | | | | | |
|---------------|---------|----|---------------|---|----------|--------|
| SPAPB17E12.08 | intron1 | 40 | SPAPB17E12.08 | N-glycosylation protein | -3.27283 | -0.678 |
| SPBC27B12.08 | intron7 | 39 | sip1 | Pof6 interacting protein Sip1 | -3.25232 | -0.06 |
| SPAC664.03 | intron1 | 37 | SPAC664.03 | RNA polymerase II associated Paf1 complex | -3.24804 | -0.152 |

Table 3.6

Top twenty down-regulated introns in the Long Stress sample vs. the Long Rich sample.

| gene | intron | intron length | gene short name | gene description | log2 intron stress/rich | log2 gene rpkm stress / rich |
|---------------|---------|---------------|-----------------|---|-------------------------|------------------------------|
| SPBC16H5.04 | intron2 | 101 | SPBC16H5.04 | pho88 family protein | 4.99136 | -0.023 |
| SPBC106.12c | intron3 | 106 | SPBC106.12c | THO complex subunit | 4.56412 | -0.296 |
| SPAC19G12.02c | intron2 | 102 | pms1 | MutL family mismatch-repair protein Pms1 | 4.43035 | 0.295 |
| SPBC32H8.08c | intron2 | 102 | omh5 | alpha-1,2-mannosyltransferase Omh5 | 4.13937 | -0.47 |
| SPBC3D6.15 | intron2 | 104 | rps2501 | 40S ribosomal protein S25 | 3.93746 | -0.356 |
| SPAC19D5.03 | intron2 | 101 | cid1 | poly(A) polymerase Cid1 | 3.90396 | -0.079 |
| SPAC13A11.04c | intron2 | 100 | ubp8 | SAGA complex ubiquitin C-terminal hydrolase Ubp8 | 3.85349 | 0.372 |
| SPAC3H5.08c | intron1 | 106 | SPAC3H5.08c | WD repeat protein, human WDR44 family | 3.71768 | -0.147 |
| SPBC13A2.01c | intron1 | 108 | SPBC13A2.01c | nuclear cap-binding complex small subunit | 3.59217 | -0.205 |
| SPBC800.04c | intron1 | 100 | rpl4301 | 60S ribosomal protein L37a | 3.34113 | -0.723 |
| SPAC31A2.02 | intron2 | 99 | trm112 | protein and tRNA methyltransferase regulatory subunit | 3.24126 | 0.044 |
| SPAC1486.02c | intron1 | 107 | dsc2 | Golgi Dsc E3 ligase complex subunit Dsc2 | 3.08408 | 0.183 |
| SPCC1682.09c | intron2 | 99 | SPCC1682.09c | mitochondrial guanine nucleotide transporter | 3.08405 | -0.115 |
| SPBC418.01c | intron1 | 106 | his4 | imidazoleglycerol-phosphate synthase | 2.94379 | 0.795 |
| SPAC5D6.05 | intron1 | 102 | sep11 | mediator complex subunit Pmc6 | 2.89002 | 0.098 |
| SPAC1782.07 | intron3 | 103 | qcr8 | ubiquinol-cytochrome-c reductase complex subunit 7 | 2.80652 | 0.969 |
| SPCC74.01 | intron2 | 101 | sly1 | SNARE binding protein Sly1 | 2.72468 | -0.155 |

| | | | | | | |
|---------------|---------|-----|---------------|---|---------|--------|
| SPAC12G12.11c | intron2 | 106 | SPAC12G12.11c | DUF544 family protein | 2.6601 | 0.34 |
| SPAC2F3.18c | intron1 | 105 | SPAC2F3.18c | ribonuclease kappa ortholog | 2.65344 | -0.232 |
| SPAC328.05 | intron4 | 98 | SPAC328.05 | RNA-binding protein involved in export of mRNAs | 2.5089 | -0.427 |

Table 3.7

Top twenty up-regulated introns in the Short Stress sample vs. the Short Rich sample.

| gene | intron | intron length | gene short name | gene description | log2 intron stress/rich | log2 gene rpkm stress / rich |
|--------------|---------|---------------|-----------------|---|-------------------------|------------------------------|
| SPCC1739.01 | intron1 | 108 | SPCC1739.01 | zf-CCCH type zinc finger protein | -3.05292 | -0.233 |
| SPBC28F2.12 | intron2 | 40 | rpb1 | RNA polymerase II large subunit Rpb1 | -2.8117 | -0.438 |
| SPCC285.16c | intron1 | 115 | msh6 | MutS protein homolog | -2.70251 | -0.609 |
| SPBC11G11.01 | intron1 | 113 | fis1 | mitochondrial fission protein Fis1 | -2.5138 | 0.53 |
| SPCC576.15c | intron1 | 109 | ksg1 | serine/threonine protein kinase Ksg1 | -2.48828 | 0.016 |
| SPCC1902.02 | intron1 | 114 | mug72 | oxidoreductase | -2.31309 | -0.356 |
| SPAC4D7.09 | intron2 | 126 | tif223 | translation initiation factor eIF2B gamma subunit | -2.25094 | -0.853 |
| SPAC4G8.07c | intron4 | 131 | SPAC4G8.07c | tRNA (m5U54) methyltransferase Trm2 | -2.148 | -0.703 |
| SPAC17G6.15c | intron3 | 113 | SPAC17G6.15c | MTC tricarboxylate transporter | -2.07524 | -1.06 |
| SPBC887.18c | intron3 | 118 | hfi1 | SAGA complex subunit Hfi1 | -2.0614 | 0.109 |
| SPAC16A10.04 | intron4 | 116 | rho4 | Rho family GTPase Rho4 | -2.035 | -0.372 |
| SPCC613.08 | intron2 | 121 | SPCC613.08 | CDK regulator, involved in ribosome export | -2.00353 | -1.472 |
| SPAPYUG7.04c | intron2 | 123 | rpb9 | DNA-directed RNA polymerase II complex subunit Rpb9 | -1.94585 | -0.365 |
| SPAC27E2.10c | intron4 | 141 | rfc3 | DNA replication factor C complex subunit Rfc3 | -1.91991 | -0.619 |
| SPAC4G9.14 | intron3 | 109 | SPAC4G9.14 | mitochondrial Mpv17/PMP22 family protein 2 | -1.91033 | -0.544 |
| SPBC3H7.15 | intron1 | 122 | hhp1 | serine/threonine protein kinase Hhp1 | -1.87402 | -0.552 |
| SPAC56F8.05c | intron1 | 155 | mug64 | BAR domain protein | -1.87082 | -0.704 |

| | | | | | | |
|-------------|---------|-----|-------------|--|----------|--------|
| SPCC1442.06 | intron3 | 137 | pre8 | 20S proteasome complex subunit alpha 2, Pre8 | -1.80873 | -0.149 |
| SPAC3G9.03 | intron1 | 111 | rpl2301 | 60S ribosomal protein L23 | -1.77462 | -1.767 |
| SPAC4D7.04c | intron1 | 112 | SPAC4D7.04c | cis-prenyltransferase | -1.77038 | -0.608 |

Table 3.8

Top twenty down-regulated introns in the Short Stress sample vs. the Short Rich sample.

3.3 Discussion

The lariat sequencing method presented here transcends the difficulties of its predecessor in profiling short lariats. The shortest *S. pombe* introns are efficiently recovered (Figures 3.1b and 3.2), allowing the discovery of over 500 novel introns. Over one quarter of these occur in previously “intronless” genes, expanding the known catalogue of genes amenable to splicing-mediated gene regulation. While independent confirmation exists for many of these (164 from the previous lariat sequencing experiment and 92 from RNA-seq junction reads (, 2011), nonetheless the results presented here would benefit from large-scale validation. Further, since the method itself would certainly gain power with further optimization of each of the steps in the sequencing library preparation, there are doubtless many introns still unannotated in the *S. pombe* genome. Based on a simple extrapolation of the fact that at least half of the sequences in this library preparation were adapter dimers and thus non-informative, a sequencing run in which this problem were resolved might be expected to yield (as an upper bound) roughly another 500 novel introns.

3.4 Materials and Methods

Strains Used

All experiments were performed using one of four strains.

1. **wild type 972 h⁻** from ATCC, # 38366

(<http://www.atcc.org/ATCCAdvancedCatalogSearch/ProductDetails/tabid/452/Default.aspx?ATCCNum=38366&Template=fungiYeast>)

2. ***Δdbr1*** DBR1::Nat, constructed using standard cloning techniques from the wild type strain above.

Growth Conditions of *Δdbr1* Cells

For the “rich” RNA sample, the cells were grown as described in the “YES vegetative growth (samples #1-#7)” part of the Materials and Methods section in Chapter 2, except that cells were grown to an OD of 0.6

For the “stress” RNA sample, a subset of the stress conditions used for the study in Chapter 2 was used (see table below). Cells were grown as described in the “Growth Conditions of *Δdbr1* Cells” part of the Materials and Methods section in Chapter 2, with the specific stresses shown in the table below.

| media | condition | temperature (°C) | OD | minutes | Stressor concentration |
|-------|-------------------------------|---------------------|-----|---------|---------------------------|
| YES | vegetative growth | 30 | 0.1 | | |
| YES | vegetative growth | 30 | 0.6 | | |
| YES | vegetative growth | 30 | 1 | | |
| YES | vegetative growth | 30 | 5 | | |
| YES | vegetative growth | 37 | 0.6 | | |
| YES | vegetative growth | 16 | 0.6 | | |
| YES | heat shock | 25-42 | 0.6 | 10 | |
| YES | heat shock | 25-42 | 0.6 | 60 | |
| YES | cold shock | 30-16 | 0.6 | 30 | |
| YES | cold shock | 30-16 | 0.6 | 180 | |
| YES | salt stress-KCl | 30 | 0.6 | 20 | 0.5M |
| YES | salt stress-KCl | 30 | 0.6 | 120 | 0.5M |
| YES | DNA dmg-MMS | 30 | 0.6 | 20 | 0.02% |
| YES | DNA dmg-MMS | 30 | 0.6 | 120 | 0.02% |
| YES | Lithium | 30 | 0.6 | 20 | 0.1 M |
| YES | Lithium | 30 | 0.6 | 120 | 0.1 M |
| YES | Ethanol | 30 | 0.6 | 20 | 10% |
| YES | Ethanol | 30 | 0.6 | 120 | 10% |
| YES | H ₂ O ₂ | 30 | 0.6 | 20 | 0.32mM |
| YES | H ₂ O ₂ | 30 | 0.6 | 120 | 0.32mM |
| EMM | 3-AT | 30 | 0.6 | 30 | 50 mM |
| EMM | 3-AT | 30 | 0.6 | 180 | 50 mM |
| EMM | no glucose | 30 | 0.6 | 20 | |
| EMM | no glucose | 30 | 0.6 | 120* | |
| EMM | no Nitrogen | 30 | 0.6 | 20 | |
| EMM | no Nitrogen | 30 | 0.6 | 120* | |
| EMM | no Phosphorus | 30 | 0.6 | 20 | |

| | | | | |
|----------|-------------------|----|-----|--------|
| EMM | no Phosphorus | 30 | 0.6 | 120* |
| YES-Agar | solid media | 30 | | 3 days |
| EMM | vegetative growth | 30 | 0.6 | |

Table 3.S1
Growth conditions used for the “stress” samples.

Total RNA Isolation from *Δdbr1* Cells

This was carried out exactly as described in the Materials and Methods section of chapter 2.

Lariat RNA isolation via Two-Dimensional PAGE

Total cellular RNA from *Δdbr1* strains grown under a variety of conditions was isolated as described above for the “rich” and “stress” samples. For each of these two samples, two different sets of gel running conditions were used in order to optimize the recovery of lariat RNAs of different lengths. To recover RNA lariats as short as ~30nt in length, “short” gel running conditions were set up as follows. For the rich sample, 65ug of RNA in 10uL of RNase free water was mixed with 10uL 2X loading buffer (90% w/v formamide, 0.05M EDTA pH8), boiled for two minutes, placed on ice and then loaded onto a single lane of a 7.5% acrylamide 8.3M urea gel in 1X TBE. The gel was run at 200V constant for 2 hours and 33 minutes. This lane was cut out of the gel and recast across the top of a 15% acrylamide 8.3M urea gel in 1X TBE. This second gel was run at 400V constant for 6 hours and 45 minutes. For the “stress” sample, the short gel was run as follows: First 100ug RNA in 10uL water and 10uL loading buffer was run on 7.5% acrylamide 8.3M urea gel in 1X

TBE at 200V constant for 2 hours and 30 minutes. The second gel was 15% acrylamide 8.3M urea gel in 1X TBE, run at 400V constant for 6 hours and 45 minutes. For both samples the gels were stained with sybr gold for 30 minutes in 1X TBE and destained for 10 minutes, and the lariat arc was excised using a new razor blade. Both pre-cut and cut gels for both the “short rich” and “short stress” gels are shown below.

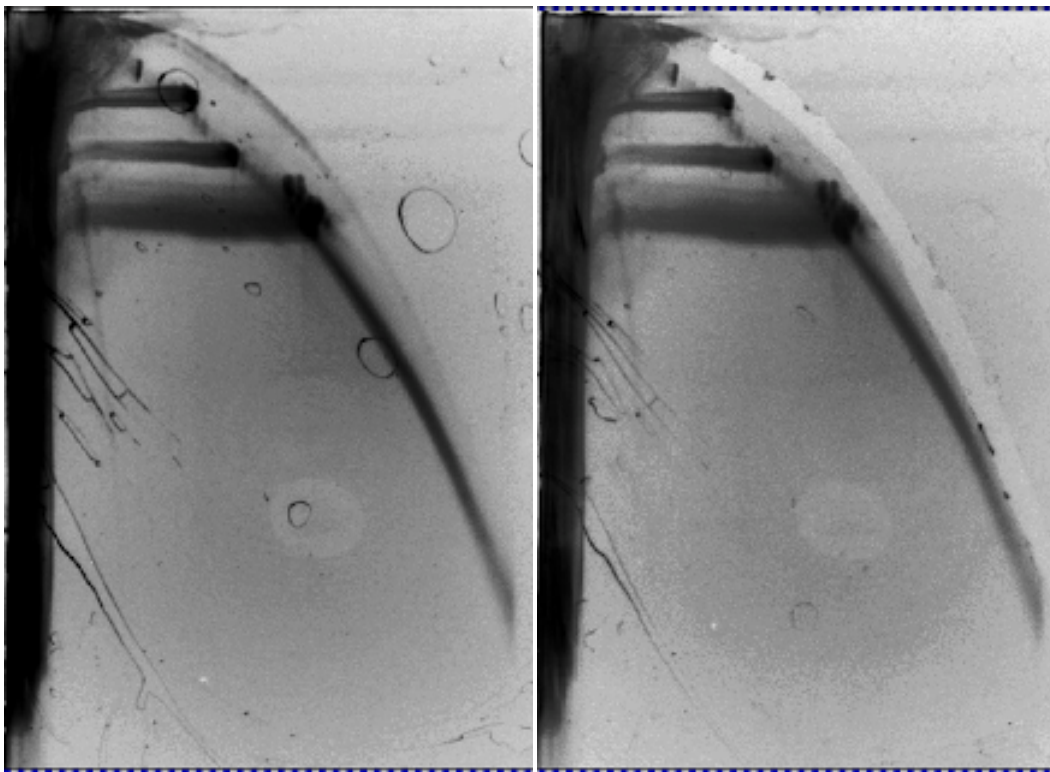


Figure 3.S1

“Short Rich” lariat get, before and after cutting the lariat arc.

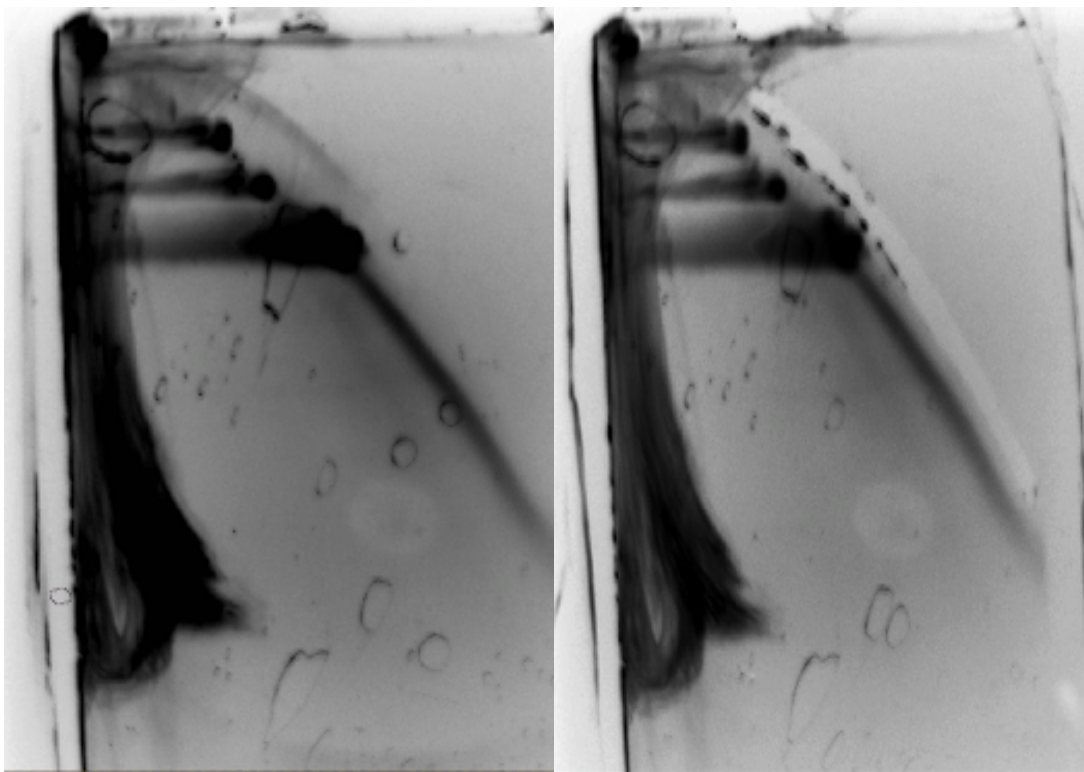


Figure 3.S2

“Short Stress” lariat get, before and after cutting the lariat arc.

To obtain longer lariats with lengths up to ~380nt, a different set of gel conditions, hereafter referred to as the “long” gel conditions, was used. For the rich RNA sample, 65ug of RNA in 10uL of RNase free water was loaded as described above onto a single lane of a 4% acrylamide 8.3M urea gel in 1X TBE. The gel was run at 200V constant for 2 hours and 33 minutes. This lane was cut out of the gel and recast across the top of a 7.5% acrylamide 8.3M urea gel in 1X TBE. This second gel was run at 400V constant for 6 hours and 45 minutes. For the “stress” sample, the long gel was run as follows: First 100ug RNA in 10uL water and 10uL loading buffer was run on 4% acrylamide 8.3M urea gel in 1X TBE at 200V constant for 2 hours and 30 minutes. The second gel was 7.5% acrylamide 8.3M urea gel in 1X TBE, run at 400V constant for 6 hours and 45 minutes. For both samples the gels were stained with sybr gold for 30 minutes and destained for 10 minutes, and the lariat arc was excised using a new razor blade. Both pre-cut and cut gels for both the “long rich” and “long stress” gels are shown below.

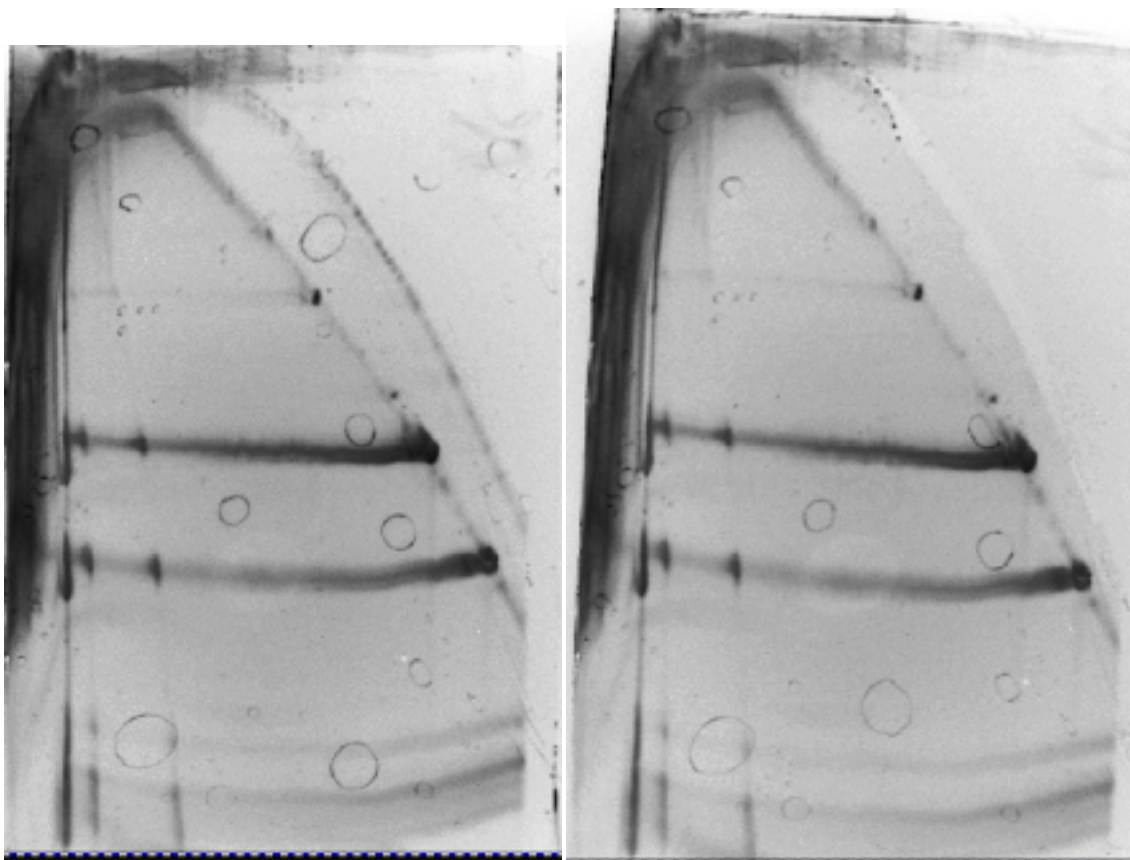


Figure 3.S3

“Long Rich” lariat get, before and after cutting the lariat arc.

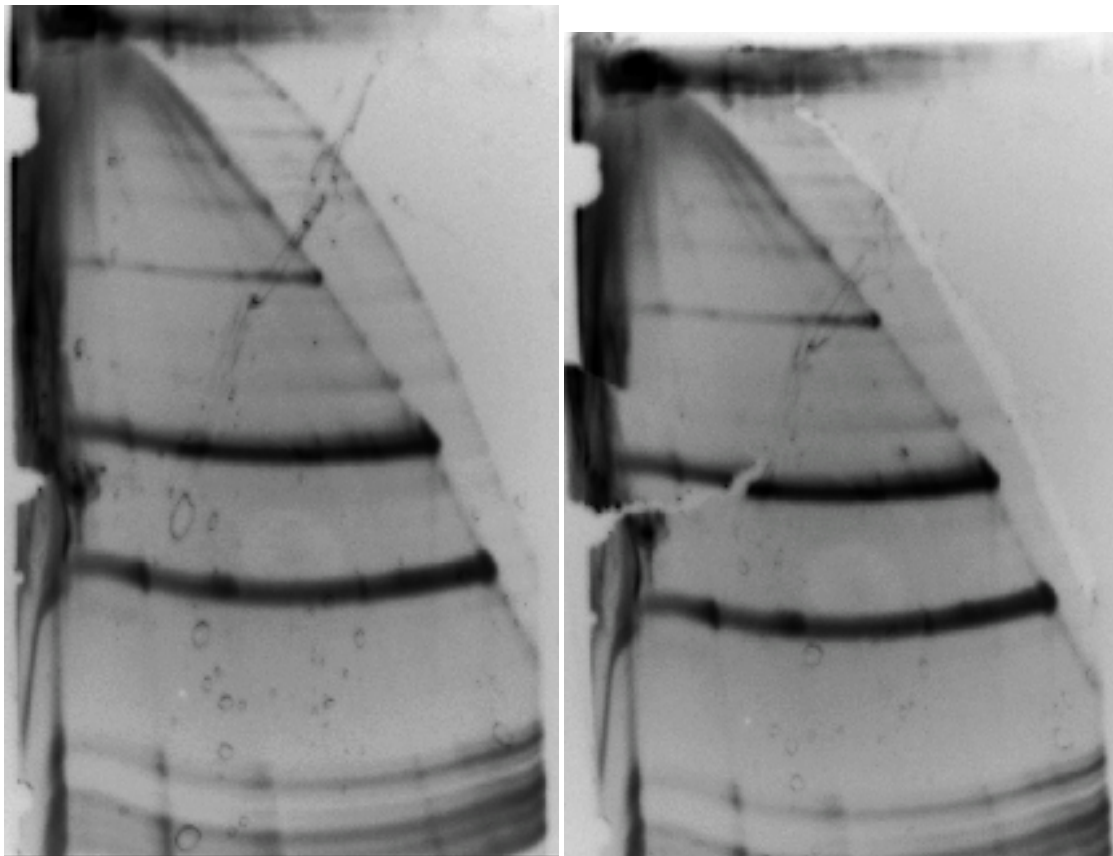


Figure 3.S4

“Long Stress” lariat get, before and after cutting the lariat arc.

The four samples from here onwards are referred to as:

“short rich” (SR)

“short stress” (SS)

“long rich” (LR)

“long stress” (LS)

I: Elution of RNA from gel arcs

The arcs from the four different gels each weighed between 1 and 2 grams. RNA was eluted from these arcs as follows. First, each arc was finely diced using a fresh razor blade into ~1mm cubes. Then each diced arc was transferred to a new 15mL falcon tube, to which 6mL 0.3M sodium acetate was added. Each tube was allowed to rotate for 21 hours at 4 degrees Celsius, and then centrifuged for 5 minutes at 4750 rpm. For each of the four samples supernatant was removed, split into two 2.5mL aliquots and transferred into two new 15mL falcon tubes. To each such tube, 7.5mL 100% ethanol was added, this was vortexed, and then each tube was left at minus 20 °C for twenty four hours.

II: Ethanol precipitation of the 0.3M sodium acetate slurry

Each of the four samples was aliquoted from the 15mL falcon tubes into 2mL tubes. All samples had 10 2mL tubes, except for LS, which only had 6. 1uL 20ug/uL glycogen was added to each 2mL tube, then all tubes were centrifuged for 30 minutes at >15,000rpm at 4 degrees. Each sample was washed twice with 2mL 70% ethanol (10 minute spins at >15,000rpm at 4 degrees), then dried under a vacuum at room temperature for 30minutes. Each resulting RNA pellet in each of the 10 tubes was resuspended in 3uL water (except for LS, for which 5uL water was used for each of the 6 tubes). These RNA solutions were pooled for each condition, and used in a DBR1 reaction as described below.

III: DBR1 reaction and precipitation

For each of the four samples, a DBR1 Reaction was set up as follows:

- 17uL RNA sample
- 2uL 10X DBR1 reaction buffer (500 mM Tris-HCl pH 7.0, 40 mM MnCl₂, 25 mM DTT, 250 mM NaCl, 0.1% Triton X-100, 1 mM EDTA, 1.5% Glycerol)
- 1uL of enzyme (generous gift from Beate Schwer)

These components were mixed and incubated at 30 degrees for 60 minutes.

Ethanol precipitated was carried out as follows. A precipitation mix was set up:

- 20 uL DBR1 reaction
- 2uL 3M NaOAc pH 5.3
- 1uL 20ug/uL glycogen
- 60uL 100% ethanol

The precipitation mix was incubated at -20 degrees °C for 30 minutes. The sample was then centrifuged at >15,000rpm at 4 °C for 30 minutes, and washed 2X with 70% ethanol (room temperature, 500uL each time), for five minutes. Finally, the sample was vacuum dried at room temperature and resuspended in 19uL water

IV: Fragmentation and precipitation

Fragmentation of the DBR1 treated RNA was performed as follows. First, a fragmentation mix was set up with:

- 18uL post-precipitation from previous step
- 2uL 10X zinc chloride buffer (100mM ZnCl₂, 100mM Tris HCl pH 8)

These components were mixed in a 200uL PCR tube, and put into thermocycler at 70 degrees for 5 minutes. The reaction was stopped by adding 2uL 0.5M EDTA, then placing on ice.

Isopropanol precipitation was carried out as follows. First, the precipitation mix was set up:

- 20uL fragmentation reaction
- 2uL 3M NaOAc pH 5.3
- 1uL 20ug/uL glycogen
- 20uL 100% isopropanol

The precipitation mix was centrifuged at >15,000rpm at 4 degrees for 20 minutes, then washed 2X with 70% ethanol (room temperature, 300uL each time) for five minutes. Finally the sample was vacuum dried at room

temperature and resuspended in 17uL water.

V: Phosphatase treatment and PNK treatment and precipitation

(adapted from the Illumina Directional mRNA seq

protocol: Directional_mRNA-Seq_SamplePrep_Guide_15018460_A.pdf)

Phosphatase treatment was performed as follows. In a 200uL PCR tube, the following was mixed:

- 16uL sample from previous step
- 2uL 10X Antarctic Phosphatase buffer (NEB)
- 1uL Antarctic phosphatase (NEB)
- 1uL RNase inhibitor (Tru-Seq small RNA kit)

This mix was incubated at 37 °C for 30 minutes, then 65°C for 5 minutes to inactivate enzyme

PNK treatment was carried out by adding the following to the 200uL PCR tube from the phosphatase treatment:

- 5uL 10X PNK buffer (NEB)
- 5uL 10mM ATP (Tru-Seq small RNA kit)
- 1uL RNase inhibitor (Tru-Seq small RNA kit)
- 2uL PNK (NEB)

- 0.5uL 0.1M DTT

- 16.5uL water

This reaction was incubated at 37°C for 60 minutes.

Isopropanol precipitation was undertaken by setting up the following precipitation mix:

- 50uL fragmentation reaction

- 5uL 3M NaOAc pH 5.3

- 1uL 20ug/uL glycogen

- 50uL 100% isopropanol

This precipitation mix was centrifuged at >15,000rpm at 4 degrees for 20 minutes, then washed 2X with 70% ethanol (room temperature, 500uL each time) for five minutes. Finally, the sample was vacuum-dried at room temperature and resuspended in 6uL water

VI: 3' and 5' adapter ligation (taken from the Illumina Tru-Seq Small

RNA cloning protocol

(TruSeq SmallRNA SamplePrep Guide 15004197 A.pdf)

3' adaptor ligation was performed as follows. The following components

(mixture A) were mixed in a 200uL PCR tube:

- 1uL 3' RNA adapter
- 5uL RNA from previous step

The mixture was incubated on pre-heated thermocycler at 70 °C for 2minutes, then moved to ice immediately. At this point a thermocycler was preheated to 28°C. The following components (mixture B) were mixed in a new 200uL PCR tube on ice (volumes are give first per reaction, then x4.4 in parentheses, which was enough for four samples):

- 2uL 5X HM Ligation buffer (8.8uL)
- 1uL RNase inhibitor (4.4uL)
- 1uL T4 RNA Ligase 2 Truncated (4.4uL)

4uL of mixture B to mixture A for each sample. This combination was mixed by pipetting 6 times and centrifuged briefly, then incubated at 28°C on the preheated thermocycler for 1 hour. 1uL stop solution was then added to the reaction, this was mixed by pipetting 6 times, and the incubation was continued

for an additional 15 minutes.

5' adaptor ligation was carried out as follows. 4.4uL 5' adaptor was added to a 200uL PCR tube (this is enough for four samples, with 1uL per sample). The tube was incubated in a pre-heated thermocycler at 70°C for 2 minutes, then immediately moved to ice, and a thermocycler was then preheated to 28°C. The following components were then added to the tube with the 5' adaptor (volumes are given first per reaction, then x4.4 in parentheses, which is enough for four samples):

- 1uL 10mM ATP (4.4uL). Pipetted 6 times and centrifuged briefly to mix.
 - 1uL T4 RNA Ligase (4.4uL). Pipetted 6 times and centrifuged briefly to mix.
- 3uL of this mixture was added to each sample tube from the post-3'adaptor ligation reaction. This reaction tube was incubated on the pre-heated thermocycler at 28°C for 1 hour, then moved to ice.

VII: Reverse transcription (taken from the Tru-Seq Small RNA cloning protocol (TruSeq_SmallRNA_SamplePrep_Guide_15004197_A.pdf) and test PCR amplification

An RT reaction of the adapter-ligated RNA was set up as follows. First, the following components (mixture A) were mixed in a 200uL PCR tube:

- 6uL 3' and 5' adapter-ligated RNA from previous step
- 1uL RT primer (RTP)

These components were pipetted up and down 6 times and centrifuged briefly.

The mixture was incubated on a pre-heated thermocycler at 70°C for 2 minutes, then the tube was immediately placed on ice. At this point a thermocycler was preheated to 50°C. The following components (mixture B) were mixed in a new 200uL PCR tube (volumes are give first per reaction, then x4.4 in parentheses, which is enough for four samples):

- 2uL 5X First Strand Buffer (8.8uL)
- 0.25uL 25mM dNTP mix (1.1uL)
- 0.25uL water (1.1uL)
- 1uL 100mM DTT (4.4uL)
- 1uL RNase inhibitor (4.4uL)
- 1uL superscript II Reverse Transcriptase (4.4uL)

These components were pipetted 6 times up and down and centrifuged briefly.

5.5uL of mixture B was added to each sample mixture A. This combination was further mixed by pipette 6 times and then centrifuged briefly. The tube was then incubated in the pre-heated thermocycler at 50°C for 1 hour.

VIII: PCR amplification, gel elution, pooling of samples

A PCR reaction was set up for each sample as follows. The following

components were added to a 200uL PCR tube for each sample:

- 30uL water
- 10uL 5X HF Phusion Buffer
- 2uL RP1
- 0.5uL 25mM dNTPs
- 0.5uL Phusion polymerase
- 2uL RP1X (index primer: 28 - SR, 33 - SS, 25 - LR, 30 - LS)
- 5 uL sample

The index primers used were 28 for the Short Rich sample, 33 for the Short Stress sample, 25 for the Long Rich sample and 30 for the Long Stress sample.

The following PCR program was run:

1: 98 degrees 30s

2: 21 cycles:

98 degrees 10s

60 degrees 30s

72 degrees 15s

The PCR reactions were run on a 6% non-denaturing polyacrylamide gel at 12W for one hour. The gel was stained with sybr gold in 1X TBE for 30

minutes and destained for 10 minutes. The gels are shown before and after cutting below.

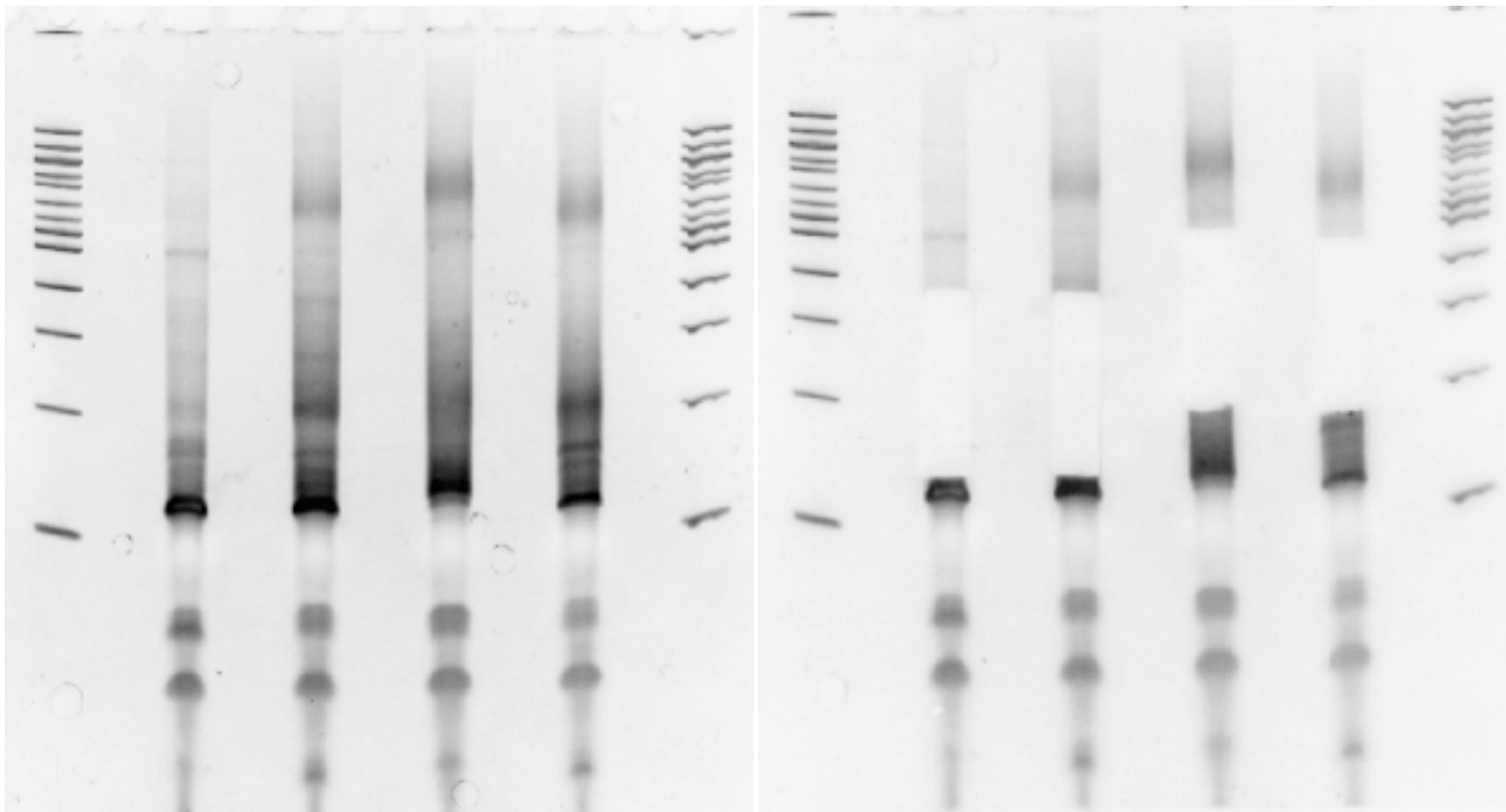


Figure 3.S5

All four sequence library samples, post PCR amplification, before (left) and after (right) cutting for elution. The lanes are from left to right: Short Stress (SS), Short Rich (SR), Long Stress (LS), Long Rich (LR).

It is not apparent from the images, but the final gel slices that were used for sequencing were cut higher up than what the cut gel shows for SS and SR. On the uncut gel, in the SR lane, 4 distinct bands can be seen above the prominent adapter band (~120bp). For the sample that was to be sequenced, the lane was cut down to include the 4th and 3rd distinct bands. However, the region corresponding to the 1st and 2nd bands was also cut and separately eluted and saved, just in case the main sample did not have the small introns.

After cutting out and razor-dicing the relevant parts of the lane for each sample, the weights of gel for each sample were found to be:

LR: 0.44g; LS: 0.45g; SR: 0.35g; SS: 0.35g

900uL 0.3M NaOAc was added to each of the diced gels in 15mL tubes and DNA was eluted from the gel slices under rotation for 10 hours at 4 °C, then for 2 hours at room temperature. The samples were then centrifuged for 5 minutes at ~4000xg, and the supernatant was split into two 400uL aliquots per sample into two 2mL tubes, adding 1.2mL 100% ethanol and 1uL 20ug/uL glycogen to each tube.

All the tubes were incubated at -20°C for two hours, and then a precipitation was performed as follows. First, the samples were centrifuged at >15,000xg for

30 minutes, then they were washed 2X with 2mL 70% ethanol. Each wash used a 10 minute centrifugation step at $>15,000 \times g$. The samples were then vacuum-dried at room temperature. Finally, the pellet in each tube was resuspended in 6uL water, and pooled into one tube for each of the four samples.

2uL out of 12uL for each of the four samples was used for a Qubit quantification, with the following results:

LR: 7.05ng/uL; LS: 6.2ng/uL; SR: 7.81ng/uL; SS: 2.21ng/uL

The samples were pooled these samples equally by taking into account the average molecular weight of the samples. By looking at the upper and lower bounds of where the gel was cut (see above) for each sample, it was found that the midpoint of each sample was as follows:

LR and LS: 315bp; SR and SS: 255bp

This method provided "average molecular weights" with which the samples could be normalized. The following normalized concentrations were obtained:

LR: 5.71ng/uL; LS: 5.02ng/uL; SR: 7.81ng/uL; SS: 2.21ng/uL

To obtain normalized concentrations, the long samples were divided by the ratio of the long to the short average molecular weight. In fact this normalization was incorrect: instead, the calculation should have been performed to equalize all four samples based on # of molecules rather than on

concentration. Ultimately, the final volumes submitted for sequencing to give a "normalized weight" of 12.5ng per sample was:

LR: 2.2uL; LS: 2.5uL; SR: 1.8uL; SS: 5.5uL

Intron Discovery

Sequence reads were aligned to the *S. pombe* reference genome from the Sanger centre website using bowtie version 0.10.1, allowing for zero mismatches, and unique alignments. After read alignment, “peaks” were defined as non-zero read regions with at least twenty-five reads. To perform the assessment of intron lengths for figure 2, the distribution of the lengths of peaks that overlapped annotated introns with a maximum difference of 10nucleotides in the positions was calculated. The annotated introns that were considered were all of those from the Sanger_Schizosaccharomyces_pombe.ASM294v1.17.gtf file obtained from pombase (www.pombase.org), plus those novel introns discovered in the previous lariat sequencing experiment (chapter 2). The intron length distribution was calculated for each of the four samples (“ShortRich”, “ShortStress”, “LongRich”, “LongStress”).

To discover novel introns, a probabilistic inference of five prime splice sites and branch points and three prime splice sites was performed for each peak with a minimum length of 20 nucleotides and a minimum read count of twenty five, as described in Chapter 2, with the following differences. First,

unlike the sequencing approach in Chapter 2, the approach used here retained strand information for the sequencing reads, so the windows around peak boundaries were only scanned in one strand. Second, the size of the window in this case was restricted to 10 nucleotides downstream of the peak 5' boundary when searching for the five prime splice site, and 15 nucleotides upstream and downstream of the peak 3' boundary when searching for the branch point and the three prime splice site. This process was carried out for all four sequencing samples.

Classifying Inferred Introns

All of the introns inferred as described above were classified into four main categories with twenty-one total sub categories, as follows.

Main categories:

“known”: The inferred intron mostly overlaps a previously annotated intron whose annotation was independent of lariat sequencing.

“larseq1”: The inferred intron mostly overlaps a previously annotated intron that was discovered in the previous lariat sequencing experiment.

“novel”: The inferred intron is completely novel , not overlapping any “known” or “larseq1” introns

“prev_intronless_novel”: The inferred intron is completely novel ,and occurs in a gene for which there were previously no annotated introns

For the “known” and “larseq1” main categories, there were four sub-categories, as follows:

single_intron: both splice sites of the inferred intron coincides with the splice sites of a known or larseq1 intron.

alt_5ss: only the three prime splice site of the inferred intron coincides with the three prime splice site of a known or larseq1 intron.

alt_3ss: only the five prime splice site of the inferred intron coincides with the three prime splice site of a known or larseq1 intron.

alt_splice_sites: neither splice sites of the inferred intron coincides with the splice sites of a known or larseq1 intron, but the inferred intron overlaps a known or larseq1 intron, and that is the only genomic feature it overlaps.

For the “novel” and “prev_intronless_novel” main categories, there were six sub-categories, as follows:

5utr_intron: the novel intron occurs in the annotated 5’utr of a protein coding gene.

3utr_intron: the novel intron occurs in the annotated 3’utr of a protein coding gene.

cds_in-frame: the novel intron occurs inside an annotated coding exon of a protein coding gene, and has a length that is a multiple of three.

cds_frameshift: the novel intron occurs inside an annotated coding exon of a protein coding gene, and has a length that is not a multiple of three.

ncRNA: the novel intron occurs inside an annotated exon of a non coding RNA gene.

antisense: the novel intron does not overlap any genomic features in it's own strand, but overlaps at least one feature in the opposite strand.

unclassified: the novel intron overlaps no features in either strand.

Finally, there are two remaining sub-categories that pertain to an inferred intron overlapping multiple features in its own strand:

exon_skip: the inferred intron shares one each of its splice site with two *different* introns within the same gene, and completely overlaps the intervening exon(s).

multi: the inferred intron overlaps multiple genomic features and does not fall into any of the other categories.

Calculations of Differential Expression for Inferred Introns

Inferred introns that were present in both a “rich” dataset and the analogous “stress” dataset (i.e. present in both “ShortRich” and “ShortStress” or in both “LongRich” and “LongStress”) were grouped and read counts were compared. To assign an inferred intron as present in both datasets, the intron in one dataset had to have splice sites with genomic locations that were no further

than 20 nucleotides away from those of the intron in the other dataset. Allowing this difference permitted a comparison of introns that almost completely overlapped each other except for a small region due to usage of alternative splice sites in one or the other condition (i.e. rich vs. stress). Read counts for both introns in such a pair were normalized by dividing by the relative sizes of the alignable sample. For example, when comparing an intron from the “LongRich” sample to the analogous intron from the “LongStress” sample, read count in the LongStress sample was divided by ~ 2.8 relative to the read counts in the LongRich sample, which was the ratio of the number of alignable reads in the LongStress sample relative to the number of alignable reads in the LongRich sample (~ 14 million to ~ 5 million). Next, log₂ differences were calculated by dividing the normalized read count from the stress sample by that in the rich sample, and taking the logarithm base two of that ratio.

To allow analysis of those introns with no analogue in one or the other sample (i.e. ran intron with reads in the stress sample but zero reads in the rich, or vice versa), pseudocounts of ‘1’ were added to every normalized read count for all introns before calculating the log₂ ratio.

Bibliography

Khalid, M. F., Damha, M. J., Shuman, S., & Schwer, B. (2005). Structure-function analysis of yeast RNA debranching enzyme (dbr1), a manganese-dependent phosphodiesterase. *Nucleic Acids Research*, 33(19), 6349-60. doi:10.1093/nar/gki934

Pratico, E. D., & Silverman, S. K. (2007). Ty1 reverse transcriptase does not read through the proposed 2',5'-branched retrotransposition intermediate in vitro. *RNA (New York, N.Y.)*, 13(9), 1528-36. doi:10.1261/rna.629607

Rhind, N., Chen, Z., Yassour, M., Thompson, D. A., Haas, B. J., Habib, N., . . . Heiman, D. I. (2011). Comparative functional genomics of the fission yeasts. *Science*, 332(6032), 930. doi:10.1126/science.1203357

Wood, V., Gwilliam, R., Rajandream, M. -A., Lyne, M., Lyne, R., Stewart, A., . . . Baker, S. (2002). The genome sequence of *Schizosaccharomyces pombe*. *Nature*, 415(6874), 871-880.

Wood, V., Harris, M. A., McDowall, M. D., Rutherford, K., Vaughan, B. W., Staines, D. M., . . . Kersey, P. J. (2012). PomBase: A comprehensive online resource for fission yeast. *Nucleic Acids Research*, 40(D1), D695-D699.

Zhang, Z., Hesselberth, J. R., & Fields, S. (2007). Genome-wide identification of spliced introns using a tiling microarray. *Genome Research*, 17(4), 503-9. doi:10.1101/gr.6049107

CHAPTER 4: MAPPING *S. POMBE* BRANCHPOINTS ON A GENOMIC SCALE

4.1 Introduction

During splicing, the 2' hydroxyl group on an adenosine residue internal to the intron nucleophilically attacks the 5' phosphodiester bond at the guanine residue of the five prime splice site, forming a lariat intron (Madhani & Guthrie, 1994). In this lariat structure, an unusual 2'-5' phosphodiester bond is formed, which requires enzymatic activity to break (Khalid, Damha, Shuman, & Schwer, 2005). The adenosine residue involved in this 2'-5' linkage and several residues immediately surrounding it are collectively termed the “branch point” (Wahl, Will, & Lührmann, 2009). The branch point sequence is known to interact with components of the splicing machinery during initial assembly and during catalysis (Staley & Guthrie, 1998; Wahl et al., 2009).

While the five prime and three prime splice sites of introns in a multitude of organisms has been characterized on genome-wide scales using high-throughput technologies (Guttman et al., 2010; Wang et al., 2008; Zahler, 2012; Zhang, Hesselberth, & Fields, 2007), the same is not true for the branch points of introns. Such information for branch points could be used to better predict binding sites for components of the splicing machinery (Kafasla et al., 2012; Wang, Zhang, Lynn, & Rymond, 2008), to make better predictions of

intron three prime splice sites when building gene models computationally (Ter-Hovhannisyan, Lomsadze, Chernoff, & Borodovsky, 2008; Zhang, Kuo, & Chen, 2012), and might even prove useful in disease contexts (Janssen et al., 2000; Khan et al., 2004).

In an RNA-seq experiment, reads that map to the exon-exon junctions created by splicing can be used to precisely demark the five and three prime boundaries of an intron (Tang & Riva, 2013; Trapnell et al., 2012). Whereas the number of such reads per intron is mathematically predicted to be proportional to read sequencing depth, gene expression level and read length, the number of RNA-seq reads mapping a branch point is predicted to be far fewer, for two reasons.

In order to precisely map a branch point using RNA-seq data, reads must traverse the 2'-5' branch structure in an intact, circular lariat (Figure 4.1b). The reads thus produced have a “back to front” order whereby the read aligns to the genome in two segments, and the order of those segments is reversed in the genome sequence relative to the read sequence (Figure 4.1c). However, lariats are typically short-lived species (Khalid et al., 2005; Nam, Lee, Trambley, Devine, & Boeke, 1997), which is one factor that would decrease their sampling in RNA-seq experiments. Second, unlike for an exon-exon junction used to map the ends of an intron, reverse transcriptase has a highly reduced rate of traversal across the 2'5'-phosphodiester linkage at the branch point

(Pratico & Silverman, 2007). Thus cDNA synthesis during sequencing library preparation will reduce the percentage of branch point-traversing reads in the cDNA relative to in the RNA.

Nonetheless, a recent study leveraged the existence of huge RNA-seq data sets in humans to find over 2,000 branch point traversing reads in over a billion RNA-seq reads, mapping the branch points for over 800 human introns (Taggart, Desimone, Shih, Filloux, & Fairbrother, 2012). To my knowledge, to date this is the only example of precise branch point mapping on a genomic scale. Since the study presented in Chapter 2 stabilizes and enriches circular lariats relative to RNA-seq in the fission yeast *Schizosaccharomyces pombe* (Figure 4.1a and see Chapter 2) and sequences without disrupting the 2'-5' phosphodiester linkage (unlike the study presented in Chapter 3), it was reasoned that this dataset could prove a useful source for branch point locations.

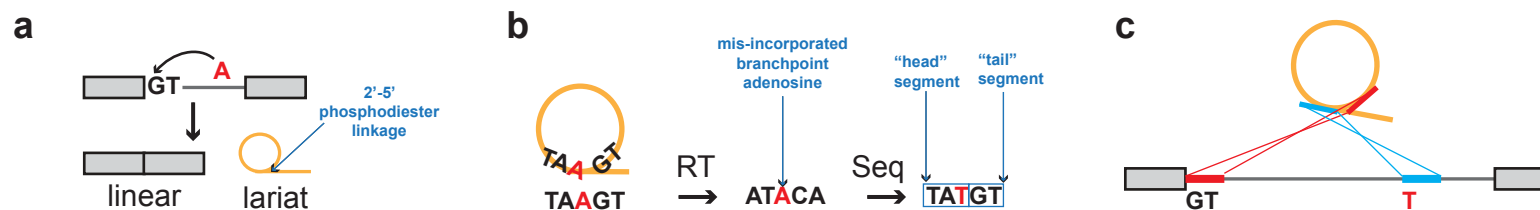


Figure 4.1

Finding Branch point traversing reads in the Chapter 2 lariat sequence data. **(a)** Splicing produces a branched lariat structure with an unusual 2'-5' phosphodiester linkage at the branch. **(b)** Reverse transcriptase often mis-incorporates an adenosine in place of a thymidine at the branch point, leading to an apparent A to T transition in sequencing data. **(c)** Branch point traversing reads align in back-to-front order within an intron, where the "tail" fragment aligns to the five prime splice site, and the "head" fragment to the branch point, with a mismatch at the branch point adenosine

4.2 Results

To find branch point-traversing reads in the dataset from Chapter 2, a search was conducted in analogous fashion to that used for the previous human study for reads that aligned as two separate segments with a length of at least 10 bases each within an *S. pombe* intron, but in back-to-front order (Figure 4.1c and see Methods). Specifically, the “tail” fragment of the read was required to align to the five prime splice site of an intron, and the “head” fragment to align downstream within the same intron.

To precisely map the branch points using such reads, the high mis-incorporation rate of reverse transcriptase through the 2'-5' phosphodiester bond at the branch point (Taggart et al., 2012) was taken advantage of. Thus, by requiring that the only mismatch within the read occurs at the terminal base of the “head” fragment (Figure 4.1b) the branch point adenosine can be precisely mapped. Such an event nearly always caused an A to T transition specifically at the branch point adenosine in the observed read sequence in this dataset. The lack of mismatches anywhere else in the read over a length of 39 bases and unique alignment of these reads in a genome with a size under 14 megabases, coupled with the back to front nature of the read segments and the consistent mutation from A specifically to T, increases confidence that these reads are *bona fide* branch point-traversing reads and not spuriously mis-aligned.

A second, rarer event observed in this dataset involved an omission by reverse transcriptase of one or two bases including the branch point adenosine. Such events also allowed precise branch point identification.

Altogether, over 24,000 branch point-spanning reads were identified in this manner, mapping the branch points for over 800 *S. pombe* introns (Table 4.1*). Considering that the lariat sequencing dataset consisted of ~15 million total reads, the detection of branch point traversing reads shows around an 800 fold enrichment in sensitivity relative to the previous human study (Taggart et al., 2012). While the number of introns sampled is comparable to that of the previous human study, this is likely related to the technical difficulties of sampling shorter introns using the lariat sequencing approach that preserves 2'-5' linkages (see Chapters 2 and 3). Indeed, 800 is over half of the total number of introns sampled in any fashion by the dataset from Chapter 2, whereas 800 represents a far smaller percentage of the human intronome sampled by the sequencing in the previous study.

Confirming Exon Skipping events from Chapter 2 using Branchpoint traversing reads

A handful of branchpoint-traversing reads contained the branchpoint of a downstream intron juxtaposed with the five prime splice site of an upstream intron. These reads represented branchpoint traversing reads for exon skipping

events. A total of 12 such reads were found, independently validating 8 exon skipping events from chapter 2 (Table 4.2)

Table 4.2

Branchpoint traversing Reads that Confirm exon-skipping events from Chapter

| gene | as exon | read count | read example |
|--------|---------|------------|---|
| pth1 | 3 | 1 | CATTTAATTTCACTACTATGTAAGTAATATCCAAAGCAAT |
| hus1 | 3 | 3 | ATTTGATTGCCCTAACTATGTAAGTGAAGATATTAAGCAT |
| sr140 | 2 | 1 | TCAGTTTAAATATGTATGTAAATATTGATTATAAGGTAA |
| alp41 | 3 | 2 | TCTTAAGCTTGTAAAGTATCTTTTAAGGCTCCATAATTTAT |
| alp41 | 5 | 1 | TTTATGAATCCAAGCTGTAGGCTGAAAATGTGAATTATCT |
| ats1 | 3 | 2 | ATATGAGCTGTGTTTCGTAAAAAAGTAACTTGTTTGAT |
| duf757 | 3 | 1 | GATGGAATAAAAAATGTACGTGGTTGTAGAAAGTATTTAG |
| pca1 | 2 | 1 | AGAACATCTTGCTATGTACGGAATGCGAAAGGTCAAGTAA |

Comparing Empirically Discovered Branch point Sequences to Computationally Inferred Ones

To understand how these empirically determined branch point sequences compared to computationally predicted ones, branch point sequence motifs were constructed as follows. First, the “empirical” set of branch points was defined as all of the branch point adenosines that were discovered for the 800+ *S. pombe* introns mentioned above. Next, branch point sequences were constructed for each of these by including the five bases upstream of the branch point adenosine and the one base downstream, making a seven base branch point sequence. Finally, these were used as input to the weblogo program to create a sequence motif logo (Crooks, Hon, Chandonia, & Brenner, 2004). An analogous procedure was applied to construct a branch point sequence motif from all of the computationally predicted branch point sequences in the pombase resource website:

(ftp://ftp.ebi.ac.uk/pub/databases/pombase/pombe/Archived_directories/Intron_Data/OLD/aligned_introns)

(Wood et al., 2012).

A comparison of the two motifs shows that purely at the level of bulk sequences, the empirically derived motif is almost identical to the computationally derived one (Figure 2). However, this bulk comparison only compares sequence similarity at a gross level. The results are confounded by

the fact that often sequences that look like branch points occur multiple times towards the three prime ends of *S. pombe* introns, such that for many of the *S. pombe* intron multiple potential branch points are predicted, without any information as to whether any of them are actually used. (ftp://ftp.ebi.ac.uk/pub/databases/pombase/pombe/Archived_directories/Intron_Data/OLD/README).

To obtain a finer-grained understanding of the ability to accurately predict branch points computationally in *S. pombe*, for every intron in which at least one empirically discovered branch point sequence existed in our dataset, the location within the intron of the empirical branch point was compared with the location of the computationally inferred branch point. There were 50 introns for which none of the predictions had any empirical support (Table 4.3) and 121 introns for which at least one prediction had no empirical support. For 6 introns with empirically discovered branch points in our dataset, no computational predictions had been made, likely due to the inexistence of a “canonical” branch point sequence within a suitable distance from the three prime splice site

(ftp://ftp.ebi.ac.uk/pub/databases/pombase/pombe/Archived_directories/Intron_Data/OLD/README).

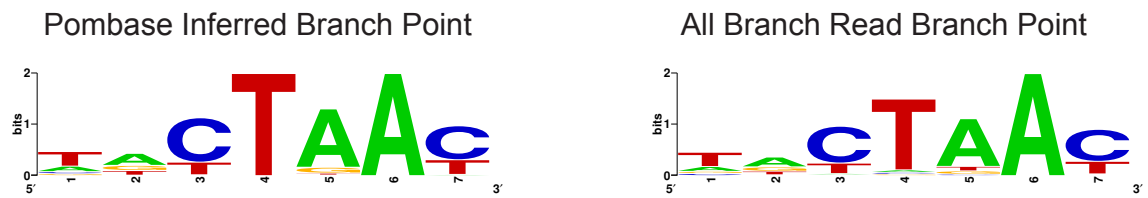


Figure 4.2

A comparison of 4642 computationally inferred branch point sequences (left) vs. 1002 empirically discovered branch point sequences (right).

Table 4.3

Introns for which empirical discovered branch points should replace computational inferences

| gene,chr,strand,f5ss,3ss | correct bp inferences | total bp inferences | empirical bp,bp-3ss-dist |
|----------------------------------|--------------------------|------------------------|-----------------------------|
| SPCC16A11.08,3,+,883703,883821 | 0 | 0 | TATTAAC,45 |
| SPAC24C9.02c,1,-,3040087,3039996 | 0 | 2 | TACTAAC,11 |
| SPAC22G7.08,1,+,749996,750112 | 0 | 1 | CGCTAAC,12 |
| SPCC24B10.14c,3,-,924780,924668 | 0 | 1 | AATTAAC,40 |
| SPBC16G5.12c,2,-,4235632,4235544 | 0 | 1 | TGCTAAC,10 |
| SPCC622.14,3,+,1427311,1427631 | 0 | 1 | CCGCTAT,192 |
| SPBC649.02,2,+,535459,535631 | 0 | 1 | TTCTAAC,10 |
| SPAC17A5.16,1,+,1788093,1788191 | 0 | 1 | TGCTAAC,13 |
| SPCC290.02,3,+,1993002,1993175 | 0 | 1 | TTTTAAT,41 |
| SPAC664.12c,1,-,1728816,1728731 | 0 | 1 | TACTAAC,10 |
| SPCC1795.08c,3,+,981129,981210 | 0 | 3 | AACTAAT,20 |
| SPBC1105.10,2,+,3524760,3524874 | 0 | 1 | GTTTAAC,10 |
| SPBC3F6.03,2,+,4189968,4190266 | 0 | 2 | AAAATAA,217 |
| SPBC12D12.03,2,+,2307687,2307930 | 0 | 1 | AATTAAT,77 |
| SPAC13G6.05c,1,-,182000,181750 | 0 | 2 | TACTAAC,104 |
| SPAC11E3.10,1,+,5300949,5301115 | 0 | 1 | ACTAAAC,13 |
| SPAC1486.02c,1,-,3188048,3187942 | 0 | 1 | TTCTAAC,15 |
| SPBC409.06,2,+,1142345,1142495 | 0 | 2 | TACTAAC,18 |
| SPAC328.10c,1,-,3497276,3496989 | 0 | 1 | ATTGTAG,195 |
| SPAC1565.04c,1,-,1296413,1296256 | 0 | 0 | TACTAAG,9 |
| SPBC1347.07,2,+,4074119,4074237 | 0 | 0 | TACTAAC,10 |

| | | | |
|-----------------------------------|---|---|---------------------------|
| SPAC23H4.04,1,-,1602680,1602594 | 0 | 1 | ACTAAAC,10 |
| SPCC1620.06c,3,-,2154999,2154848 | 0 | 2 | AACTAAC,17 |
| SPAC630.06c,1,-,356983,356905 | 0 | 1 | TATT ^T TAT,13 |
| SPAC1399.05c,1,+,1838853,1838969 | 0 | 2 | ACTAAAC,12 GTAATAC,41 |
| SPAC959.09c,1,-,3403135,3403050 | 0 | 1 | TATTAAC,12 |
| SPAC15A10.08,1,+,3693163,3693302 | 0 | 2 | ATACTAA,11 ACTAAAT,9 |
| SPAC30D11.13,1,-,1095141,1094921 | 0 | 1 | AGCTAAT,59 |
| SPCC63.07,3,+,849232,849327 | 0 | 0 | AACTAAC,32 |
| SPAC17G8.04c,1,-,2348704,2348615 | 0 | 0 | CACTAAT,12 |
| SPAC1687.15,1,+,930197,930677 | 0 | 1 | CGTCAAT,362 |
| SPAC17A5.07c,1,-,1764876,1764793 | 0 | 1 | TACTAAG,10 |
| SPCC18.11c,3,-,1976347,1976264 | 0 | 1 | TGCTAAC,11 |
| SPCC1183.03c,3,-,596666,596513 | 0 | 1 | AGTT ^T TAG,32 |
| SPAC11H11.03c,1,-,4780022,4779668 | 0 | 1 | ATCTAAT,226 |
| SPBC19G7.17,2,+,2382853,2383027 | 0 | 2 | TGCTAAT,10 |
| SPBC21C3.05,2,+,3805928,3806243 | 0 | 1 | CGCTAAC,172 |
| SPBC800.08,2,+,265469,265604 | 0 | 2 | ATCT ^T TAC,18 |
| SPBC776.15c,2,-,3208458,3208167 | 0 | 1 | T ^T TGCTAA,195 |
| SPAC17A5.07c,1,-,1766336,1766189 | 0 | 0 | AACTAAC,33 |
| SPAC1006.03c,1,-,5073388,5073189 | 0 | 1 | AATACAT,37 |
| SPBC9B6.11c,2,-,1834578,1834450 | 0 | 1 | TATTAAC,10 |
| SPAC1705.02,1,+,1568870,1569063 | 0 | 1 | TGCT ^T TAT,91 |
| SPCC285.17,3,+,1832743,1832913 | 0 | 2 | AGCTAAG,16 |
| SPAC31G5.17c,1,-,3020229,3020011 | 0 | 1 | ACAATAT,96 |
| SPCC1393.09c,3,-,816393,816248 | 0 | 1 | GATT ^T TCAA,62 |

| | | | |
|----------------------------------|---|---|---|
| SPBC1709.16c,2,-,1131978,1131824 | 0 | 2 | GAATAAC,14 |
| SPAC23H4.03c,1,+,1604397,1604529 | 0 | 1 | T [*] TCTGAC,16 |
| SPBC8D2.01,2,+,1358926,1359482 | 0 | 1 | T [*] T [*] T [*] TAAT,421 |
| SPAC1F7.04,1,+,4224251,4224581 | 0 | 1 | GGT [*] T [*] TAC,246 |

Using Branch Points to Compare Polypyrimidine Tracts in Introns in non Coding RNA Regions vs. Protein coding regions

In multi-cellular eukaryotes with introns that are typically many times longer than exons, it has long been known that RNA-binding proteins bound to sequences in exonic regions of protein coding genes help to recruit the splicing machinery at the upstream three prime splice site and downstream five prime splice site (Smith & Valcárcel, 2000; Wu & Maniatis, 1993).

In recent years it has become increasingly apparent that epigenetic signals such as nucleosome positioning (Schwartz, Meshorer, & Ast, 2009; Tilgner et al., 2009) and DNA methylation (Shukla et al., 2011) from protein-coding exons as well as differential GC percentage within exons vs. introns (Amit et al., 2012) serve to demark protein coding exons from introns. However, it is less clear how the splicing machinery recognizes introns that separate non-coding exons (referred herein as “introns in non-coding RNA regions”), such as those within untranslated regions or in noncoding RNA genes (Eden & Brunak, 2004). Compared with protein coding exons and the introns that separate them, the difference in epigenetic marks and base composition between non-coding exons and introns in non-coding RNA regions is less pronounced (, 2004). Relative to the five prime splice site, the three prime splice site sequence itself is relatively information poor (Lim &

Burge, 2001), but in addition to the branch point sequence, an additional sequence element called the “polypyrimidine tract” has been shown to play an important role in recruitment of splicing factors to the three prime end of the intron (Staley & Guthrie, 1998; Wahl et al., 2009). The polypyrimidine tract is located between the branch point sequence and the AG dinucleotide of the three prime splice site, and, as its name suggests, is relatively rich in pyrimidine bases.

In a model in which introns in non-coding regions are more dependent than introns separating protein-coding exons on the polypyrimidine tract for proper identification of the 3' splice site, one simple prediction is that polypyrimidine tracts are stronger in the former class of introns than in the latter. To test this idea, all introns for which a branch point sequence was empirically identified were separated into two classes: introns from non coding RNA regions, and introns separating protein coding exons, based on annotation (see Methods). First, the branch point sequences themselves were compared, with motifs constructed as described above. These sequences did not differ significantly (Figure 4.3a). Next, for all introns in both classes, the polypyrimidine tract was defined as the bases between the branch point adenosine and the A of the AG dinucleotide of the three prime splice site. For each polypyrimidine tract, the percentage of T and C residues was calculated. Though the number of introns in both classes differed greatly – 36 vs. 967 –

there is over counting of introns with multiple branch points), the difference in mean pyrimidine percentage of the polypyrimidine tract was great enough to achieve strong statistical significance despite the low sample size of the non-coding RNA intron class (56.96 vs. 50.07 $p < 0.0032$, one-sided t-test, Table 4.4). Importantly, the difference in pyrimidine percentage for the intronic sequence comprising the twenty bases upstream of the branch point was not significantly different between the two classes (58.95 vs. 57.55 $P < 0.2864$, one-sided t-test, Table 4.4). This result was also tested by constructing sequence motifs for the terminal seven nucleotides upstream of the polypyrimidine tract for both classes of intron (Fig 4.3b).

Table 4.4

Pyrimidine percentage comparisons for polypyrimidine tract and upstream intronic sequence in introns from non-coding RNA regions (ncRNA introns) vs. those separating protein coding exons (pce introns)

| Sequence Type | ncRNA Intron Mean | pce Intron Mean | p-value |
|----------------------|-------------------|-----------------|-----------------|
| polypyrimidine tract | 56.96 | 50.07 | 0.003175 |
| upstream sequence | 58.95 | 57.55 | 0.2863 |

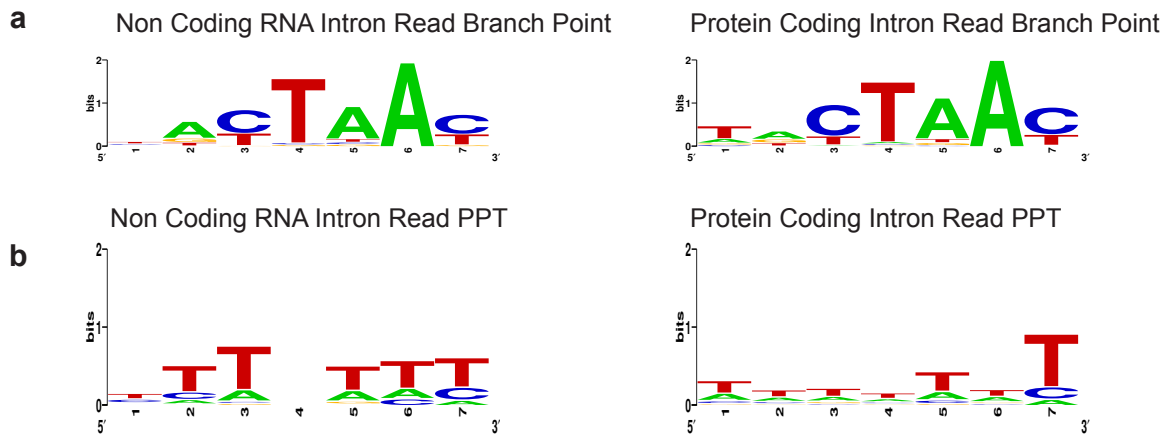


Figure 4.3

A comparison of sequence motifs for intronic elements in introns from non-coding RNA regions vs. those separating protein coding exons. **(a)** branch point sequence motifs. **(b)** polypyrimidine tract terminal sequence motifs.

4.3 Methods

Detecting branch point-spanning reads

To detect branch points, the same approach as outlined by Taggart *et al* (Taggart et al., 2012) to search for inverted alignments within single reads, with slight modification as follows. All reads from the lariat sequencing dataset from Chapter 2 were aligned to the *S. pombe* genome using bowtie (Langmead, 2010). Only those reads that could not be aligned with fewer than three mismatches were considered further. The remaining reads were each split into all possible pairs of two fragments (a “head” and a “tail” fragment) with a minimum length of 10 bases for each fragment. These pairs of fragments were then aligned to the *S. pombe* genome using bowtie, as follows.

To account for the frequent substitution rate of reverse transcriptase traversing the 2-5' phosphodiester linkage between the five prime splice site and branch point in a lariat (Taggart et al., 2012) these head and tail fragments were aligned using the bowtie “soft clip” options -5 1 and -3 1 (which do not consider the terminal bases during alignment), along with the zero mismatch and unique alignment options (-v 0 - m1). This meant that a single substitution at the terminal end of one of the fragments would still produce an alignment. The reason both ends of both fragments were soft clipped is because the

original data set was not strand specific, so the read sequence obtained could have been transcribed from either strand.

For all pairs of fragments for which both fragments aligned to the genome, branch point-spanning reads for annotated introns were searched for in the following manner. Only pairs where both fragments aligned within the same gene were considered further. Conceptually, for a major spliceosomal branch point-spanning read, the tail fragment starts with a GT (as it represents the 5' splice site), and its upstream coordinate in the intron strand will match that of the five prime splice site. The head fragment will align downstream of the tail fragment in the intron strand, but upstream of the intron's three prime splice site (for a canonical splicing event). Since the read information was not strand-specific, however, each read for which both fragments aligned within a gene were considered as potential "cognate" reads (i.e. the read sequence is the same as that of the intron) or "reverse" reads (i.e. the read sequence is the reverse complement of the intron sequence). Thus, in addition to searching for reads where the tail fragment begins with "GT" and aligns upstream of the tail fragment, with its upstream co-ordinate matching that of the intron five prime splice site, a search was also conducted for reads where the *head* fragment *ended* in "AC", and the tail fragment aligned downstream of the head fragment in the

intron's strand. Such a scenario would be true for the reverse complement of a branch point-spanning read.

For a “cognate” branch-point spanning read that represents an exon-skipping event, the head fragment will align in a downstream intron, upstream of that intron's three prime splice site. A search was conducted for both “cognate” and “reverse” exon-skipping branch point spanning reads. In conducting all of these searches, only previously annotated introns and introns discovered in the first lariat sequencing experiment (see Chapter 2) were considered.

To precisely map the branch point adenosine in a branch point spanning read, only reads for which the branch point sequence was mis-incorporated by reverse transcriptase were considered. Thus, for “cognate” reads, the head fragment had to have a substitution of either a “T” where the branch point “A” should have been (see Figure 4.1), or a deletion of the branch point “A”. Such mismatches allowed precise mapping of the branch point adenosine.

Classifying branch points of introns in protein coding regions vs. in noncoding RNA regions

To classify a branch point as belonging to an intron in a protein coding region, the intron containing the branch point had to separate two protein coding exons. If a branch point came from any other type of intron (untranslated regions and non-coding RNAs), it was considered a branch point from a non coding RNA region.

Comparing the polypyrimidine tract in branch points from protein coding regions vs. those from non coding RNA regions

For all branch points identified in this study, the nearest three prime splice site was calculated as follows. Aside from the annotated three prime splice site, all “AG” dinucleotides between the branch point and the annotated three prime splice site were tested for being used in splicing by searching for RNA-seq junction reads that used those potential three prime splice sites and the annotated five prime splice site. The closest such AG dinucleotide was chosen as the three prime splice site to use when determining the polypyrimidine tract. If there were no AG dinucleotides between the branch point and the annotated three prime splice site, then the annotated three prime splice site was used. To

test the statistical significance of the difference between the polypyrimidine tract percentages of the two classes of intron, the statistical pack R was used (<http://www.r-project.org/>), with the `t.test` function, and the “alternative” parameter set to “greater” to use the one-sided test for whether or not the pyrimidine percentage was greater in the introns from non-coding RNA regions vs. those that separate protein coding exons.

Constructing Sequence Motifs for Figures 4.2 and 4.3

Branch point sequences for the empirically discovered branch points were constructed as described above. For the computationally inferred branch points from pombase, branch point sequences were extracted from the input file.

(ftp://ftp.ebi.ac.uk/pub/databases/pombase/pombe/Archived_directories/Intron_Data/OLD/aligned_introns)

For polypyrimidine tracts, the last seven nucleotides of the sequence between the branch point adenosine and the A of the three prime splice site AG dinucleotide was extracted as the polypyrimidine tract. Only introns where the polypyrimidine tract was at least seven nucleotides long were considered.

For each motif, all sequences comprising the motif were entered as a file into the weblogo program website (<http://weblogo.berkeley.edu/>).

Comparing the empirically derived branch points in this study to the previously inferred branch points for the *S. pombe* genome

The locations of empirically discovered branch points were discovered as described above. Computationally inferred branch points were downloaded from the pombase website:

(ftp://ftp.ebi.ac.uk/pub/databases/pombase/pombe/Archived_directories/Intron_Data/OLD/aligned_introns/).

For each intron for which there was at least one empirically discovered branch point, the location was compared to the locations of each computationally predicted branch point. Every such try was considered an “attempt” and every match was considered a “success” here a computational prediction had empirical support. The total number of successes and number of attempts is shown in table 4.4 for introns for which at least one “failure” where a computational prediction had no empirical support.

Bibliography

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., . . .

Postolsky, B. (2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Reports*. doi:10.1016/j.celrep.2012.03.013

Crooks, G. E., Hon, G., Chandonia, J. -M., & Brenner, S. E. (2004). WebLogo: A sequence logo generator. *Genome Research*, 14(6), 1188-1190. Retrieved from Google Scholar.

Eden, E., & Brunak, S. (2004). Analysis and recognition of 5 UTR intron splice sites in human pre-mrna. *Nucleic Acids Research*, 32(3), 1131-1142. Retrieved from Google Scholar.

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., . . . Nusbaum, C. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature Biotechnology*, 28(5), 503-510.

Janssen, R., Wevers, R. A., Häussler, M., Luyten, J., STEENBERGEN-SPANJERS, G. C. H., Hoffmann, G. F., . . . HEUVEL, L. (2000). A branch site mutation leading to aberrant splicing of the human tyrosine hydroxylase

gene in a child with a severe extrapyramidal movement disorder. *Annals of Human Genetics*, 64(5), 375-382.

Kafasla, P., Mickleburgh, I., Llorian, M., Coelho, M., Gooding, C., Cherny, D., . . . Eperon, I. (2012). Defining the roles and interactions of PTB. *Biochemical Society Transactions*, 40(4), 815.

Khalid, M. F., Damha, M. J., Shuman, S., & Schwer, B. (2005). Structure-function analysis of yeast RNA debranching enzyme (dbr1), a manganese-dependent phosphodiesterase. *Nucleic Acids Research*, 33(19), 6349-60. doi:10.1093/nar/gki934

Khan, S. G., Metin, A., Gozukara, E., Inui, H., Shahlavi, T., Muniz-Medina, V., . . . Schneider, T. D. (2004). Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: Mutations result in reduced XPC mrna levels that correlate with cancer risk. *Human Molecular Genetics*, 13(3), 343-352.

Langmead, B. (2010). Aligning short sequencing reads with bowtie. *Current Protocols in Bioinformatics*, 11-7. Retrieved from Google Scholar.

Lim, L. P., & Burge, C. B. (2001). A computational analysis of sequence features involved in recognition of short introns. *Proceedings of the National Academy of Sciences*, 98(20), 11193-11198. Retrieved from Google Scholar.

Madhani, H. D., & Guthrie, C. (1994). Dynamic RNA-RNA interactions in the spliceosome. *Annual Review of Genetics*, 28, 1-26.
doi:10.1146/annurev.ge.28.120194.000245

Nam, K., Lee, G., Trambley, J., Devine, S. E., & Boeke, J. D. (1997). Severe growth defect in a *Schizosaccharomyces pombe* mutant defective in intron lariat degradation. *Molecular and Cellular Biology*, 17(2), 809-818.

Pratico, E. D., & Silverman, S. K. (2007). Ty1 reverse transcriptase does not read through the proposed 2',5'-branched retrotransposition intermediate in vitro. *RNA (New York, N.Y.)*, 13(9), 1528-36. doi:10.1261/rna.629607

Schwartz, S., Meshorer, E., & Ast, G. (2009). Chromatin organization marks exon-intron structure. *Nature Structural & Molecular Biology*, 16(9), 990-5.
doi:10.1038/nsmb.1659

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., . . . Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, 479(7371), 74-9.
doi:10.1038/nature10442

Smith, C. W., & Valcárcel, J. (2000). Alternative pre-mrna splicing: The logic of combinatorial control. *Trends in Biochemical Sciences*, 25(8), 381-387. Retrieved from Google Scholar.

Staley, J. P., & Guthrie, C. (1998). Mechanical devices of the spliceosome: Motors, clocks, springs, and things. *Cell*, 92(3), 315-26.

Taggart, A. J., Desimone, A. M., Shih, J. S., Filloux, M. E., & Fairbrother, W. G. (2012). Large-scale mapping of branchpoints in human pre-mrna transcripts in vivo. *Nature Structural & Molecular Biology*. doi:10.1038/nsmb.2327

Tang, S., & Riva, A. (2013). PASTA: Splice junction identification from rna-sequencing data. *BMC Bioinformatics*, 14(1), 116.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., & Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Research*, 18(12), 1979-1990.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., & Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. *Nature Structural & Molecular Biology*, 16(9), 996-1001. Retrieved from Google Scholar.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., . . . Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature Protocols*, 7(3), 562-578.

Wahl, M. C., Will, C. L., & Lührmann, R. (2009). The spliceosome: Design principles of a dynamic RNP machine. *Cell*, 136(4), 701-18. doi:10.1016/j.cell.2009.02.009

Wang, E. T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., . . . Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221), 470-476.

Wang, Q., Zhang, L., Lynn, B., & Rymond, B. C. (2008). A bbp--mud2p heterodimer mediates branchpoint recognition and influences splicing substrate abundance in budding yeast. *Nucleic Acids Research*, 36(8), 2787-2798.

Wood, V., Harris, M. A., McDowall, M. D., Rutherford, K., Vaughan, B. W., Staines, D. M., . . . Kersey, P. J. (2012). PomBase: A comprehensive online resource for fission yeast. *Nucleic Acids Research*, 40(D1), D695-D699. Retrieved from Google Scholar.

Wu, J. Y., & Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell*, 75(6), 1061-1070. Retrieved from Google Scholar.

Zahler, A. M. (2012). Pre-mRNA splicing and its regulation in *caenorhabditis elegans*. *WormBook*, 4(03).

Zhang, J., Kuo, C. -C. J., & Chen, L. (2012). VERSE: A varying effect regression for splicing elements discovery. *Journal of Computational Biology*, 19(6), 855-865.

Zhang, Z., Hesselberth, J. R., & Fields, S. (2007). Genome-wide identification of spliced introns using a tiling microarray. *Genome Research*, 17(4), 503-9. doi:10.1101/gr.6049107

CHAPTER 5: FUTURE DIRECTIONS

5.1 Srrm1 and Exon Definition

One of the first steps in splicing involves the recognition of five prime splice site and the three prime splice site of an intron by the U1 and U2 components of the splicing machinery, and the subsequent pairing of these components (Madhani & Guthrie, 1994; Staley & Guthrie, 1998). In unicellular eukaryotes, where introns are typically constrained to a upper length limit and are shorter than exons, U1 and U2 have been shown to interact via a protein bridge across the intron, in a model that has been termed “intron definition” (Ast, 2004; Kim, Goren, & Ast, 2008).

In multi-cellular eukaryotes however, introns are typically much longer than exons, in some cases reaching lengths of over 100kb (Keren, Lev-Maor, & Ast, 2010). This fact, coupled with the observation that internal exon lengths are typically constrained to a maximum size limit (Fox-Walsh et al., 2005; Sterner, Carlo, & Berget, 1996) led to the development of the “exon definition” model, which posits initial assembly of the U1 and U2 splicing components across an exon rather than across an intron (Berget, 1995; Robberson, Cote, & Berget, 1990). This model involves U1 snRNP binding at a five prime splice site and aiding in the recruitment of U2 snRNP at the three prime splice site of

the *upstream* intron, sometimes using factors bound on the intervening exon (Figure 5.1a) (Fox-Walsh et al., 2005; Keren et al., 2010).

While identification of many of the factors involved in intron definition has been achieved (Shao, Kim, Cao, Xu, & Query, 2012; Xu et al., 2004), less is known about the factors involved in exon definition. Genetic dissection of exon definition in a tractable model organism would likely identify more factors. The fission yeast *Schizosaccharomyces pombe*, has been proposed as just such a genetically tractable system (Ram & Ast, 2007; Webb, Romfo, van Heeckeren, & Wise, 2005), however the evidence has been mixed as to whether any of its genes undergo exon definition rather than intron definition (Romfo, Alvarez, van Heeckeren, Webb, & Wise, 2000).

Under these two different models of initial spliceosomal assembly the effects of suboptimal five prime splice site recognition due to a weak five prime splice site are expected to have different outcomes (Berget, 1995; Robberson et al., 1990). Under intron definition, the inability of the U1 snRNP to recognize the five prime splice site will lead to failure to recruit the U2 snRNP at the downstream three prime splice site, and therefore intron retention (Figure 5.1b). Under the exon skipping model by contrast, the inability of the U1 snRNP to recognize the five prime splice site will lead to failure to recruit the U2 snRNP at the *upstream* three prime splice site, and therefore exon skipping (Figure 5.1b).

With this in mind, it has been reasoned that if exon skipping is observed in conjunction with a suboptimal five prime splice site immediately downstream of the skipped exon, that the exon definition model is more parsimonious for that exon than the intron definition model (Kim et al., 2008). Therefore, the discovery of an exon-skipping event with a suboptimal five prime splice site in *S. pombe* (see Chapter 2) provides a candidate for exon

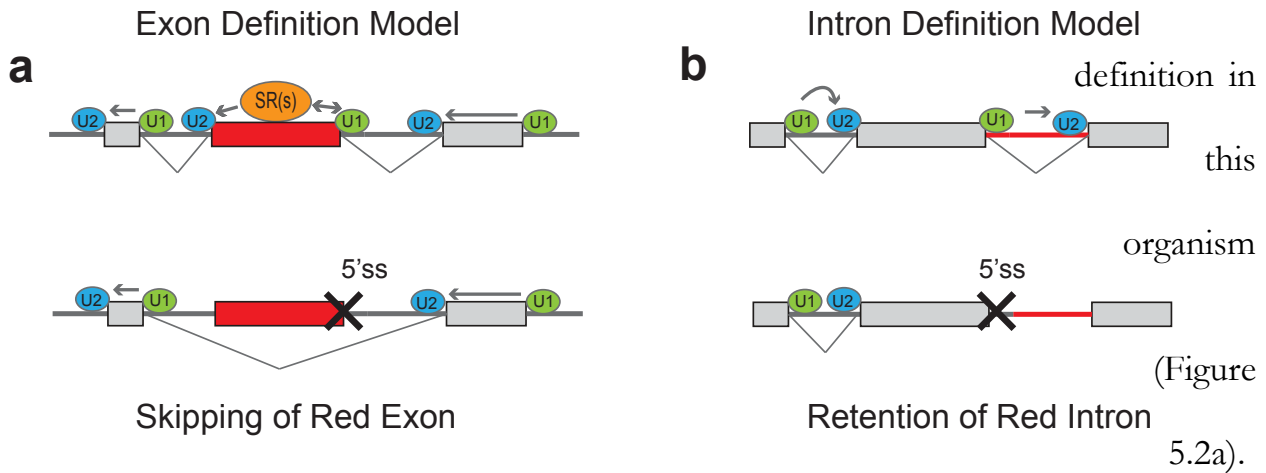


Figure 5.1

Predictions of the outcomes of failure to recruit U1 snRNP to the five prime splice site (5'ss) under the two models for initial spliceosome assembly. **(a)** Under the exon definition model, failure to recruit U1 snRNP at a 5'ss leads to failure to recruit U2 snRNP to the upstream three prime splice site (3'ss) and exon skipping. **(b)** Under the intron definition model, failure to recruit U1 snRNP at a 5'ss leads to failure to recruit U2 snRNP to the downstream 3'ss and intron retention.

A testable prediction of the exon definition model is that a mutation of an intronic five prime splice site from GT to any other dinucleotide will cause skipping of the exon immediately upstream (Berget, 1995; Robberson et al., 1990). Such a mutation will prevent the U1 snRNP from interacting at the five prime splice site, which, under the exon definition model, will lead to a failure to recruit the U2 snRNP at the upstream three prime splice site. This eventually leads to pairing of the upstream U1 snRNP to the downstream U2 snRNP and exon skipping as the internal exon fails to get defined by the splicing machinery (Figure 5.2b). By contrast, under the intron definition model, such a five prime splice site mutation would lead to retention of the intron in which it occurred (Figure 5.2c).

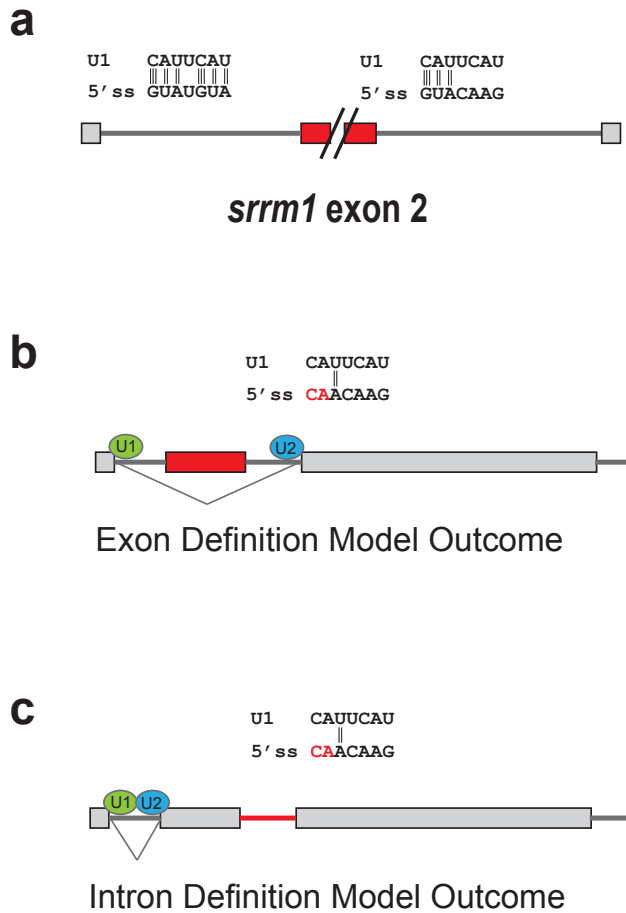


Figure 5.2

srrm1 intron 2 has a suboptimal five prime splice site (5'ss), and ablating this 5'ss leads to different outcomes under the two models for initial spliceosomal assembly. **(a)** In contrast to the 5'ss of *srrm1* intron 1, the 5'ss of intron 2 has far less homology with the RNA component of U1 snRNP, and far less hydrogen bonding by base pairing. **(b)** The expected outcome of ablating the 5'ss of intron 2 is skipping of exon 2 under the exon definition model. **(c)** The expected outcome of ablating the 5'ss of intron 2 is retention of intron 2 under the intron definition model.

In order to test whether the second exon of the *srrm1* gene undergoes exon definition, a mutant version of this gene was constructed in which the five prime splice site of intron 2 was altered from GT to CA. This version of the gene was transformed on a plasmid into a wild type *S. pombe* strain. A control strain was constructed where everything was identical except that the GT was left unmutated. These two strains were then tested for levels of intron retention and exon skipping using PCR with a forward primer in exon 1 and a reverse primer in exon 3 (Figure 5.3a).

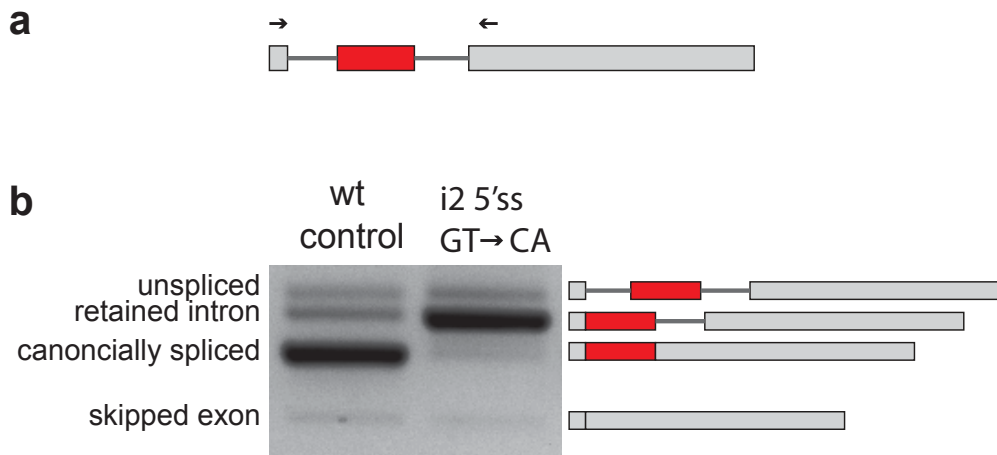


Figure 5.3

The phenotype of ablating the five prime splice site of *srrm1* intron 2 (i2 5'ss) is increased intron retention. **(a)** Schematic of primers used to amplify the different splicing isoforms around exon 2, with a forward primer in exon 1 and a reverse primer in exon 3, were the same as those used in Chapter 2. **(b)** PCR products were run on an acrylamide gel. i2 5'ss = strain with plasmid borne *srrm1* with intron 2 five prime splice site mutated from GT to CA, in a wild type background. wt control = control strain with plasmid borne wild type version of *srrm1* in a wild type background. Cartoons to the right of gel indicate identity of bands, as does the text on the left of the gel.

The results of this experiment show a striking increase in the amount of intron retention in the mutant vs. the control (Figure 5.3b), strongly suggesting that the intron definition model is at work for the *srrm1* gene. However, there are several caveats to this conclusion. First, due to technical difficulties during cloning, the upstream region of this gene that was cloned is shorter than the annotated five prime UTR (Rhind et al., 2011; Wood et al., 2012). If the annotation is correct, then the cognate promoter of this gene is not included in the cloned version. It is known that promoters can drive recruitment of specific splicing factors (Moldón et al., 2008), and if any such factors are required for exon definition, this would affect the results of the experiment. A second caveat concerns the fact that this experiment was done using plasmid-borne versions of the *srrm1* gene rather than chromosomally-integrated versions. If epigenetic factors such as chromatin context or histone modifications played a role in the exon definition of this gene, they would not apply to the plasmid-borne versions of the gene, and thus this experiment would not be able to detect exon definition. In contrast to the first caveat (short upstream region) this problem is more challenging to address, since it would involve precise point mutations at a single locus, which is difficult to achieve using homologous recombination without double selection, which could disrupt regulatory regions important for exon definition. Nonetheless, techniques exist to achieve precise, scarless chromosomal integration in the budding yeast,

Saccharomyces cerevisiae, (Storici & Resnick, 2006; Storici, Lewis, & Resnick, 2001) and these could be applied in this situation. Finally, the observation that intron retention increases upon mutation of the five prime splice site does not preclude a concomitant increase in exon skipping. Such a result would be the prediction of a model under which intron definition was the primary mode of splice site assembly for the introns flanking exon 2 of *srrm1*, but nonetheless some stabilizing interactions between splice sites occurred across exon 2. Though the exon skipped isoform was hardly visible on the gel (Figure 5.3), a more sensitive test, such as isoform-specific qPCR, Northern blotting or poisoned primer extension might be able to detect small increases in exon skipping in the mutant strain compared to the control.

5.2 Circular Exonic RNAs

Several recent studies have described a novel class of RNAs called “circRNAs”, wherein protein coding exons become circularized, precisely at their five prime and three prime ends as defined by the splice sites that flank these exons (Memczak et al., 2013). While these circRNAs have been very recently shown to play important roles in antagonizing microRNAs (Hansen et al., 2013; Memczak et al., 2013) and are present in archaea as well as eukarya (Danan, Schwartz, Edelheit, & Sorek, 2012), extremely little is known about their biogenesis (Memczak et al., 2013).

Using an approach analogous to that outlined in Chapter 3 for discovery of circular branchpoint traversing reads, I discovered reads supporting ten exonic circRNAs in the lariat sequencing data from Chapter 2 (Table 5.1)

Table 5.1

Reads from the lariat sequencing experiment in Chapter 2 that representing circRNA species. Each species is represented by a single read, except for that from cct8, which has two reads. In the last column, a truncated version of the exon sequence is shown: part of the sequence from the five prime end (the first base is the five prime end of the exon) is shown first, then “...” indicating intervening sequence, then finally part of the sequence from the three prime end of the exon is shown (the last base is the three prime end of the exon). In cases where two exons are involved, the upstream sequence is that of the first, exon, and the downstream sequence is that of the second d exon. Blue letters represent sequence from the “tail” fragment of the read, aligning perfectly to the five prime end of the exon sequence, and red letter represent sequence from the “head” fragment of the read, aligning perfectly to the three prime end of the exon sequence. Thus the reads align in “back-to front” order, indicating a circularization event.

| gene | short name | strand | circle contains | exon sequences |
|---------------|------------|--------|--------------------------|---|
| SPBC3D6.09 | dpb4 | + | exon 2 | TTCTGGGGAAATTGCTACGAATAATAACA ... CGTTAAAAAACATCTCGAAG |
| SPBC354.06 | mrps16 | + | exon 2 | GAAAGCACCTCAAGCCAAACCTATCG ... ACCGTCCGCTCTTTAGCTGAAAAG |
| SPBC337.05c | cct8 | - | exon 2, intron 2, exon 3 | ACCCGTACCTCTTTGGGTCCAAATGGA ... CTACCCAACAACAGGAGAATGAG |
| SPAC6F6.15 | ypt5 | + | exon 4, intron 4, exon 5 | CTGCTTTTTTGACTCAAACCCTTC ... TTATAAATCTCTAGCTCCTATGTACT |
| SPBC31F10.06c | sar1 | - | exon 2 | AATGACCGTTTAGCTGTAATGCAACC ... GATTGGGAGGTCATCAACAGG |
| SPBP35G2.05c | cki2 | - | exon 3 | CTGGCATTCCTAATGTTTACTAT... ACGGTGGCTATGACGGCGAAACAGATG |
| SPCC1672.01 | | + | exon 2 | ATTTTGTTTGCATGCCAGGGCAAA ... AGTTGAAGACTTATATCCCGAAGAA |
| SPBC11B10.09 | cdc2 | + | exon 2 | GAACCTATGGCGTTGTTTATAAAGCAAGAC ... AATCGATCAAATTGTGTTCG |
| SPAC6F6.07c | rps13 | - | exon 2, intron 2, exon 3 | GTCAAAAGATCATGCGTATCTTGAA ... GATCTCTACAATCTTATTAAGAAG |
| SPAC22G7.02 | kap111 | + | exon 5 | AATCATTTATGAACAATATCGACATTACAG ... CAGCAGCCGGCCAATTTATT |

To validate that these reads represent true circRNAs, I plan to use a PCR based strategy (Figure 5.4). For every circRNA exon, primers can be designed within the exon. One set of primers will be designed such that the three prime ends of the primers face each other; they are expected to amplify both cDNA derived from the linear mRNA exon and cDNA derived from the circRNA (Figure 5.4a). A second set will be designed such that the three prime ends face outwards from one another; these are expected to amplify cDNA derived from the circRNA species, but not cDNA derived from the linear mRNA exon (Figure 5.4b). Finally, a third test will be performed whereby the sample will first be treated with RNase R, an RNA nuclease that can use linear RNA but not circular RNA as a substrate (, 2013), then PCR will be carried out using the first set of primers on cDNA derived from the sample. Such a test is expected to give a markedly reduced signal for cDNA derived from the linear mRNA exon, but expected to leave the signal from cDNA derived from the circRNA species unattenuated (Figure 5.4c).

After validation, a qPCR screen can be carried out using the primers from Figure 5.4b against the entire *S. pombe* deletion strain library (Kim et al., 2010). Our lab has developed and successfully implemented such a screen in *S. cerevisiae* (Albulescu et al., 2012) and in *S. pombe* (unpublished data). Such a screen for factors affecting the abundance of circRNAs would likely prove useful for better understanding their biogenesis.

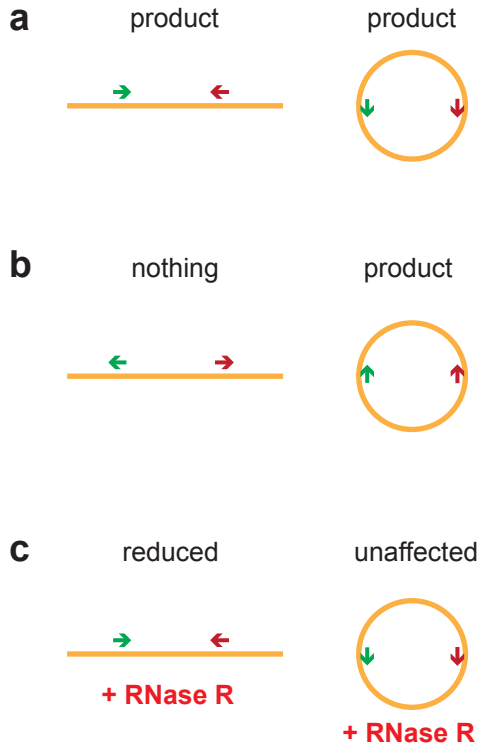


Figure 5.4

Strategy for validating circRNAs **(a)** Primers aligning within an exon whose 3' ends face towards each other are expected to amplify cDNA derived from linear exonic species as well as circular exonic species. **(b)** Primers aligning within an exon whose 3' ends face away from each other are expected to amplify cDNA derived from circular exonic species but not from linear exonic species. **(c)** RNase R treatment is expected to selectively degrade linear, not circular RNA, and thus to attenuate the PCR signal with the primer pair used in (a) for cDNA derived the linear exonic mRNA, but not for cDNA derived from circRNAs.

Bibliography

- Albulescu, L. -O., Sabet, N., Gudipati, M., Stepankiw, N., Bergman, Z. J., Huffaker, T. C., & Pleiss, J. A. (2012). A quantitative, high-throughput reverse genetic screen reveals novel connections between pre-mrna splicing and 5 and 3 end transcript determinants. *PLoS Genetics*, 8(3), e1002530. Retrieved from Google Scholar.
- Ast, G. (2004). How did alternative splicing evolve? *Nature Reviews. Genetics*, 5(10), 773-82. doi:10.1038/nrg1451
- Berget, S. M. (1995). Exon recognition in vertebrate splicing. *Journal of Biological Chemistry*, 270(6), 2411-2414. Retrieved from Google Scholar.
- Danan, M., Schwartz, S., Edelheit, S., & Sorek, R. (2012). Transcriptome-wide discovery of circular rnas in archaea. *Nucleic Acids Research*, 40(7), 3131-42. doi:10.1093/nar/gkr1009
- Fox-Walsh, K. L., Dou, Y., Lam, B. J., Hung, S. P., Baldi, P. F., & Hertel, K. J. (2005). The architecture of pre-mrnas affects mechanisms of splice-site pairing. *Proceedings of the National Academy of Sciences of the United States of America*, 102(45), 16176-81. doi:10.1073/pnas.0508489102
- Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., & Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature*, 495(7441), 384-8. doi:10.1038/nature11993

Keren, H., Lev-Maor, G., & Ast, G. (2010). Alternative splicing and evolution: Diversification, exon definition and function. *Nature Reviews. Genetics*, 11(5), 345-55. doi:10.1038/nrg2776

Kim, D. -U., Hayles, J., Kim, D., Wood, V., Park, H. -O., Won, M., . . . Palmer, G. (2010). Analysis of a genome-wide set of gene deletions in the fission yeast *schizosaccharomyces pombe*. *Nature Biotechnology*, 28(6), 617-623. Retrieved from Google Scholar.

Kim, E., Goren, A., & Ast, G. (2008). Alternative splicing: Current perspectives. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, 30(1), 38-47. doi:10.1002/bies.20692

Madhani, H. D., & Guthrie, C. (1994). Dynamic RNA-RNA interactions in the spliceosome. *Annual Review of Genetics*, 28, 1-26. doi:10.1146/annurev.ge.28.120194.000245

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., . . .

Rajewsky, N. (2013). Circular rnas are a large class of animal rnas with regulatory potency. *Nature*, 495(7441), 333-8. doi:10.1038/nature11928

Moldón, A., Malapeira, J., Gabrielli, N., Gogol, M., Gómez-Escoda, B.,

Ivanova, T., . . . Ayté, J. (2008). Promoter-driven splicing regulation in fission yeast. *Nature*, 455(7215), 997-1000. Retrieved from Google Scholar.

Ram, O., & Ast, G. (2007). SR proteins: A foot on the exon before the transition from intron to exon definition. *Trends in Genetics*, 23(1), 5-7. Retrieved from Google Scholar.

Rhind, N., Chen, Z., Yassour, M., Thompson, D. A., Haas, B. J., Habib, N., . . .

Heiman, D. I. (2011). Comparative functional genomics of the fission yeasts.

Science, 332(6032), 930. doi:10.1126/science.1203357

Robberson, B. L., Cote, G. J., & Berget, S. M. (1990). Exon definition may facilitate splice site selection in rnas with multiple exons. *Molecular and Cellular Biology*, 10(1), 84-94.

Romfo, C. M., Alvarez, C. J., van Heeckeren, W. J., Webb, C. J., & Wise, J. A.

(2000). Evidence for splice site pairing via intron definition in

schizosaccharomyces pombe. *Molecular and Cellular Biology*, 20(21), 7955-7970.

Retrieved from Google Scholar.

Shao, W., Kim, H. S., Cao, Y., Xu, Y. Z., & Query, C. C. (2012). A U1-U2

snrnp interaction network during intron definition. *Molecular and Cellular Biology*,

32(2), 470-8. doi:10.1128/MCB.06234-11

Staley, J. P., & Guthrie, C. (1998). Mechanical devices of the spliceosome:

Motors, clocks, springs, and things. *Cell*, 92(3), 315-26.

Sterner, D. A., Carlo, T., & Berget, S. M. (1996). Architectural limits on split

genes. *Proceedings of the National Academy of Sciences of the United States of America*,

93(26), 15081-5.

Storici, F., & Resnick, M. A. (2006). The delitto perfetto approach to in vivo site-directed mutagenesis and chromosome rearrangements with synthetic oligonucleotides in yeast. In Judith L. Campbell & P. Modrich (Eds.), *Methods in Enzymology: DNA repair, part B* (Vol. 409, pp. 329 - 345). Academic Press.

doi:10.1016/S0076-6879(05)09019-1

Storici, F., Lewis, L. K., & Resnick, M. A. (2001). *Nature Biotechnology*, 19(8), 773-776. doi:10.1038/90837

Webb, C. J., Romfo, C. M., van Heeckeren, W. J., & Wise, J. A. (2005). Exonic splicing enhancers in fission yeast: Functional conservation demonstrates an early evolutionary origin. *Genes & Development*, 19(2), 242-54.

doi:10.1101/gad.1265905

Wood, V., Harris, M. A., McDowall, M. D., Rutherford, K., Vaughan, B. W., Staines, D. M., . . . Kersey, P. J. (2012). PomBase: A comprehensive online resource for fission yeast. *Nucleic Acids Research*, 40(D1), D695-D699. Retrieved from Google Scholar.

Xu, Y. Z., Newnham, C. M., Kameoka, S., Huang, T., Konarska, M. M., & Query, C. C. (2004). Prp5 bridges U1 and U2 snrnps and enables stable U2 snrnp association with intron RNA. *EMBO J*, 23(2), 376-85.

doi:10.1038/sj.emboj.7600050